UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

EXPLORING THE CONSTRUCT VALIDITY AND RELIABILITY OF THE ENGLISH COMPREHENSION TEST

by

DANILLE ELIZE ARENDSE

Submitted in fulfilment of the requirements of the degree

PhD (Psychology)

in the
Department of Psychology

at the
Faculty of Humanities
University of Pretoria

Supervisor:
Prof D.J.F. Maree

Submitted:
2017

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

DECLARATION

I declare that this thesis, **Exploring the construct validity and reliability of the English Comprehension Test**, which I hereby submit for the degree PhD in Psychology at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature:____________________________

Date: _______________________________

ACKNOWLEDGEMENTS

“I am sure of this, that he who began a good work in you will bring it to completion..."

Philippians 1: 6

Dear God, I am in awe of you and your greatness. I would not have been able to achieve this without you. You have blessed this journey and I cannot thank you enough for your unending grace and mercy. I am nothing without you. Thank you for always surrounding me. Amen

I would like to acknowledge the following people who played a fundamental role in my life and who were instrumental in the completion of this dissertation.

To my mother, Magdalene Arendse, I thank you for your support, encouragement and continuous belief in me. I dedicate this paper to you, as the dream you envisioned for me has finally come to completion. I love you very much and may God continue to bless you.

To my grandfather, JPH Lewin, my sisters, Zelda and Maxine, my brother, Daniel, thank you for your love and kindness, I appreciate you. Your valuable encouragement has kept me strong and humbled me.

To my partner, Kyle Bester, thank you for your love and support, in often a difficult time. I truly appreciate all you have done for me and continue to do for me. You have made me smile and laugh amidst the chaos, I am grateful to have you in my life.

To my colleagues: Lt Col Hartzenberg, who believed in the English Comprehension Test and gave me the confidence and support I needed to pursue my studies; Lt Col Bielfeld, who has been a motherly figure and has always shown me kindness and support; Maj Machimana, my friend and confidant, I cannot thank you enough for all you are as a person and all you have done for me as a friend, you have been my source of wisdom; Maj Maine, thank you for being a friend and always supporting me; Lt Col Bruwer, Maj Chazen, Maj Cloete, Maj Ebrahim, I thank you for creating opportunities for the piloting of the English Comprehension Test and being good friends who always supported my efforts. To all of you, I say thank you and I appreciate your kindness and support.

To my supervisor, Prof David Maree, thank you for your guidance and support. Your positive feedback throughout this process has been invaluable to me. I thank you for your academic wisdom; it has made this journey unforgettable.

To the National Institute of Humanities and Social Sciences (NIHSS) SAHUDA, I would like to thank you for your financial assistance in the completion of this dissertation. Thank you for creating opportunities of scholarship, it has been crucial to the production of this dissertation. A special thanks to Prof Siphamandla Zondi, who has humbled me with his academic mentoring.

To my friends and family, whose names I cannot all mention, I would like to express my gratitude in the role you have played in my life and your support with my studies.

ABSTRACT

The consideration of a multi-cultural and multi-lingual context makes developing a test not only a challenge but a worthwhile venture. The empirically designed English Comprehension Test (ECT) is theorized to be measuring Verbal Reasoning and thus the construct validity and reliability needed to be established. The ECT was initially developed with the intention of screening for English comprehension, but this has since shifted to creating a screening tool for Verbal Reasoning. The ECT consisted of two test versions which had sample sizes of 597 for the ECT version 1.2 and 882 for the ECT version 1.3. This quantitative study aimed to ascertain whether the construct of Verbal Reasoning was measured by the ECT and in doing so, it required the use of five specific objectives. Firstly, the unidimensionality of items was explored through Rasch analyses. The results revealed that the performance of the persons was consistently misfitting across the two test version which caused deviation from the model because they were either careless, guessing or not paying attention to the questions and thus irregular answering patterns occurred. The item performance however improved across the two test versions and only contained a few items which caused differential performance which was either due to difficulty or issues related to the item content. Additionally, Rasch analyses allowed the dimensionality of the ECT to be assessed which indicated that the test was multidimensional and there was some redundancy in the items and persons which limited the variance explained.

Secondly, Confirmatory Factor Analysis (CFA) was used to confirm the dimensionality of the ECT by conducting Structural Equation Modelling. The model created was guided by the Exploratory Factor Analyses conducted on the ECT version 1.2 and performed on the data of the ECT version 1.3. The model demonstrated that the construct of Verbal Reasoning was subdivided into factors of Reasoning, Education, Vocabulary,

Deduction and Plurals. The results of the SEM indicated that the model was a good fit and the factors were indeed related to the underlying construct of Verbal Reasoning. It also indicated that the strongest factors were Reasoning and Education. This served to indicate that the different factors of the model were acceptable as it explained the construct of Verbal Reasoning.

Thirdly, supportive evidence of construct validity was attainted by conducting a Multi-trait Multi-method (MTMM) analysis. This analysis did not conform to the traditional aspects of a MTMM analysis due to the fact that some information was not available. This is however a limitation of using secondary data. These results indicated that the strongest correlations with the ECT were observed amoung the following constructs: Reading Comprehension, Vocabulary, Verbal Reasoning and long-term Memory. Moreover, the ECT was correlated with constructs that were not hypothesized to relate, such as Calculations, Mathematical Comprehension, Mechanical Insight, Spatial 2D and Spatial 3D. This emphasized that Verbal Reasoning was observed in the correlations across all constructs.

Fourthly, measurement invariance was explored using Differential Test Functioning which focused on the comparison of gender and racial groups. The DTF results for the gender comparison revealed that there were a few items which were possibly biased across the genders. The DTF results for the different racial groups (African and White & African and Coloured) indicated more possibly biased items across the racial groups. The items that were identified are a cause of concern and require further investigation. It is however worth noting that the majority of the items were considered to be appropriate for both gender and the different racial groups.

Fifthly, the internal consistency was evaluated by the Kuder-Richardson Formula 20. This reliability coefficient indicated that both test versions were reliable enough for research purposes. When exploring the items in the item-total correlations, the items which were not

adding value to the reliability coefficient were removed. This improved the internal consistency of both test versions and indicated that these test versions can be used for assessing aptitude. This is important as this test is intended for aptitude and therefore shows great promise. It should however be noted that the items which posed a threat to the reliability of the test need to be revised as they are not adding value to the internal consistency of the ECT.

All the results were then interpreted using Messick's unified theory of Construct Validity by specifically applying Messick's six facets of Construct Validity (Content, Substantive, Structural, Generalizability, External and Consequential). Based on this discussion, it was therefore concluded that the English Comprehension Test was measuring Verbal Reasoning.

The most important limitations of the study involved technical issues such as restriction of range and statistical issues which involved the limits of secondary data analyses on the statistical analyses of data. This also includes the limited exploration of Verbal Reasoning as the construct of the ECT, as all the aspects of this construct cannot be measured by the ECT. The recommendations made involve further studies to explore the possible bias observed in the items as well as additional factor analyses to explore the dimensionality. Further studies on the construct of Verbal Reasoning are also required.

The philosophical discussion on Verbal Reasoning and problematising the traditional notion of Verbal Reasoning as a Euro-American construct was significant to this study. It allowed for a new discourse on the psychological construct of Verbal Reasoning, whereby a new system of thought was created. This new discourse framed the construct of Verbal Reasoning in the ECT as both deconstructed and decolonized. This multi-faceted construct still taps into the same theoretical constructs as the traditional notion of Verbal Reasoning but avoids the use of analogies which have been observed as problematic, especially in the South African population.

Keywords: Verbal Reasoning, Construct Validity, Dimensionality, Rasch analysis, Structural Equation Modelling, Multi-Trait Multi-Method, Differential Test Functioning, Reliability, Messick, Test Development, Cognitive Testing & Psychometrics.

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAT	Academic aptitudes tests
ACM	Additive conjoint measurement
AGFI	Adjusted goodness of fit
AIC	Akaike’s information criterion
CAIC	Bozdogan’s consistent version of Akaike’s information criterion
CAT	Cognitive ability test
CFA	Confirmatory factor analysis
CFI	Comparative fit index
CHC	Carrol-Horn-Cattell
CLQT	Cognitive linguistic quick test
CTT	Classical test theory
DAF	Dispersion accounted for
DAT	Differential aptitude test
DIF	Differential item functioning
DTF	Differential test functioning
ECT	English Comprehension Test
ECVI	Expected cross-validation index
EEA	Employment Equity Act
EEOC	Equal Employment Opportunity Commission
EFA	Exploratory factor analysis
GFI	Goodness of fit
HPCSA	Health Professions Council of South Africa
HSRC	Human Sciences Research Council
ICC	Item characteristic curve
IFI	Incremental fit index
IIF	Item information function
IRT	Item response theory
ITC	International Test Commission

KZN	Kwa-Zulu Natal
LPCAT	Learning Potential Computer Adaptive Tests
MCAS	Massachusetts Comprehensive Assessment System
MNSQ	Mean-Square Statistics
MPI	Military Psychological Institute
MTMM	Multi-trait multi-method
NCP	Non-centrality parameter
NFI	Normed fit index
PAF	Principal Axis factoring
PCA	Principal Component analysis
PDSS	Postpartum depression screening scale
PGFI	Parsimony goodness of fit
PLM	Parameter logistic models
PNFI	Parsimony-normed fit index
RFI	Relative fit index
RMR	Root mean square residual
RMSEA	Root mean square error of approximation
RSMEA	Root square error of approximation
SAT	Senior Aptitude Test
SDA	Secondary data analysis
SEM	Structural equation modelling
TCC	Test characteristic curve
TIF	Test information function
TLI	Tucker-Lewis index
UNISA	University of South Africa
WAIS	Wechsler Adult Intelligence Scale
WISC	Wechsler Intelligence Scale for Children
WMLS	Woodcock Munoz Language Survey
ZSTD	Standardized residual (Z-score) Statistics

CHAPTER 1: INTRODUCTION

1.1 Aims of the Study

Language is often regarded as the most important moderator of test performance. This is because performance on assessment measures could be the product of language difficulties and not ability factors especially if a measure is administered in a language other than the test taker's home language (Nell, 2000).

Moreover, numerous researchers concur that academic success is significantly predicted by precision in language use, particularly precise command of vocabulary rather than syntactic competence or social fluency (Kilfoil, 1999; Saville-Troike, 1984). Thus, the need to assess an individual's language ability is crucial, as this would be an indicator of their academic potential. When measuring language, the concept of language has to be rendered in a measurable form. The measurable form of language is essentially the operationalizing of language in a method that makes testing possible (Auer & Wei, 2007).

This measurable form is interpreted in several ways when analysed for meaning and usage. The method in which language is analysed is also dependent on the discipline in which the language assessment is taking place (Weir, 2005). Linguists tend to focus on the different language structures found in a test, educators consider the impact of schooling and an individual's progress, sociologists would draw attention to cultural and societal influences within testing and language, while psychologists focus on the cognitive and psychological phenomena prevalent in language testing and psychometric assessment (Manktelow & Chung, 2004). Although these disciplines appear contrasting, they provide for a comprehensive examination of language testing. For this reason, language assessment intersects disciplines and requires interdisciplinary knowledge to be sufficiently contextualised.

Essentially, measurement begins with a clear description of the construct to be measured and once this theoretical objective has been achieved, the instrument is designed which will operationalize this construct. In the Psychosocial Sciences, this instrument would take the form of a test or questionnaire. The subsequent steps are validation of the instrument, which interactively informs the theoretical work around the construct (Dunne, Long, Craig & Venter, 2012; Long, 2011; Wright & Stone, 1979).

This study aspires to explore the aspect of measuring language as a psychological phenomenon, which is predominantly located within the psychological field. However, the intersections between the above-mentioned disciplines (linguistics, education, and sociology) compel one to consider these dynamics in relation to the psychological phenomenon being studied. For this reason, these various aspects will be acknowledged to assess the multifaceted reality of psychometric testing in a multicultural South Africa.

The psychometric milieu is understood in terms of the societal usage of tests and their ability to serve the needs of test users. With this in mind, the use of psychometric tests in South Africa increases annually along with the different reasons for which they are used, such as educational criteria, university admission, and employment. The psychometric tests used are, however, mostly European or American and are not specifically designed for the South African population. In addition, tests are adapted from these continents for South African use, yet some constructs are not easily adapted as there is no equivalent in the African languages.

In light of this, the construction of tests for the South African population seems to be an almost obligatory and judicious option. This option would allow South Africans to further their expertise regarding test development to bridge the gap between South Africans and those abroad. By developing tests locally, not only are candidates' performances in tests improved, but also how well the candidates understand the constructs being measured. The study, therefore, focuses heavily on test development, and as a result, the psychological construct being measured in the English Comprehension Test needs to be critically explored in both a statistical and theoretical way.

The hypothesised psychological phenomenon within the English Comprehension Test (ECT) is verbal reasoning and for this reason, academic sources have been consulted to establish if this construct is both relevant and observable in the analysis. This crucial information obtained by exploring the evidence of verbal reasoning in the test is instrumental in developing the test further and informs how the test may be utilised. The initial intention of the ECT was that it could be used to screen for English comprehension, but this focus has shifted to creating a screening tool for Verbal Reasoning.

As a result, this study advances the research and exploration of the concept of verbal reasoning, specifically within a South African population. Moreover, this study will be proven to be necessary to advance psychological research on verbal reasoning and test development in psychology. The unique nature of this study makes it indispensable to psychology in South Africa and in international psychological research. The contribution this study makes to the body of cross-disciplinary knowledge will be demonstrated as a form of social justice and intellectual advancement of locally produced measures of addressing inequality in a psychological capacity.

1.2 Contextualizing the Research

The history of South African psychological testing has been influenced by South Africa’s traumatic past. This past was characterised by two significant movements. The first movement was the formation of locally formulated languages among slaves and South African inhabitants, of which one was Afrikaans. Secondly, the influence of the Apartheid regime caused the dominance of White individuals over Black individuals. During this period, two language groups arose, namely Afrikaans and English (spoken predominately by White individuals at the time). This is significant as the influence of language was instrumental in continuing discrimination of one against another. Although racial differences were evident and initiated the racist movement of foreigners towards native Africans, the influence of languages significantly contributed to this movement. This was witnessed by foreigners concluding that native Africans had primitive communication methods and thus no sophisticated language in comparison to theirs. Another instance of this language conflict was observed in the war between the English (British) and Afrikaans (Afrikaners) speaking White individuals in South Africa. This demonstrated that language had a more profound effect than race, as being racially similar was not enough to unite the two groups. Although these White individuals eventually united in domination over Black individuals, it is worth noting that language has historically had a powerful impact in South Africa (Laher & Cockcroft, 2013).

The historical discrimination of Black individuals was then transferred into the psychological assessment of Black individuals. This form of assessment served to justify the racial discrimination prevalent in South Africa. This also allowed racism to legitimise inequality and permitted the perception to prevail that Black individuals were inferior, regardless of the lack of education or resources they received at the time. Tests were therefore used to exclude Black people from employment opportunities and created the perception of ignorant Black individuals and intelligent White individuals. This misuse of psychological tests to unfairly discriminate against Black individuals caused Black individuals to view tests negatively, which is how many still view psychometrics. This view is predominately what plagues the perception of psychometrics in South Africa, as psychometrics is continuously viewed in light of its historical past of discrimination and exclusion (Laher & Cockcroft, 2013). This is why the validation and creation of psychometrics needs to cater to South Africa’s multicultural and multilingual environment, even though it is an enormous task. This process also needs to be transparent to shift the negative perception that Black individuals have of psychometrics in South Africa.

The development of psychological measures should always endeavour to produce reliable results that are accurate when determining an individual's functioning on various constructs (Mushquash & Bova, 2007; Van de Vijver & Tanzer, 2004). This becomes difficult when dealing with individuals from different cultures, as some tests are standardised on a particular group and would therefore not be suited for use by other groups (Mushquash & Bova, 2007; Van de Vijver & Tanzer, 2004). However, it should be noted that a test can never be without bias, as some groups may still be advantaged in some way. Bias is defined as the presence of nuisance factors affecting the test scores of different groups differentially (Hambleton, Merenda, & Spielberger, 2005; Mushquash & Bova, 2007; Van de Vijver & Rothmann, 2004; Van de Vijver & Tanzer, 2004). This bias should, therefore, be reduced as far as possible, especially in cross-cultural settings (Van de Vijver & Rothmann, 2004; Van de Vijver & Tanzer, 2004).

When testing individuals from different racial or cultural groups, there are factors outside of the test which may influence the performance of individuals. These factors are often individual's exposure to education, language, test-wiseness experiences, and society, which impact on how individuals complete tests. In South Africa, language is known for its possible bias towards individuals who are not familiar with the language of the test. Consequently, in South Africa, English tests are always considered biased when individuals from the other 10 national languages complete them (Van de Vijver & Rothmann, 2004).

The strategies that have been used to address cultural bias in South Africa were implemented using translation by test developers. This translation strategy is, however, filled with its own problems as it highlights the difficulty of translating tests across 11 different languages. A noteworthy consideration is that some concepts exist in the English language, while there is no equivalent in any of the nine African languages (Schaap, 2011).

The practice of psychological testing in South Africa gave rise to institutions such as the National Institute of Personnel Research, the Interdepartmental Committee on Native Education, and the Institute for Psychological and Edumetric Research. These institutions evaluated test use in South Africa, attempted to address issues relating to unfair test use, and acknowledged the historical discrimination of Black individuals through psychometrics. These institutions led to the formation of the Human Sciences Research Council (HSRC) and the General Adaptability Battery, as a united form of addressing discrimination in psychological testing. There was a specific department that focused on testing and as time progressed, this test development department closed and the newly created tests for South Africa were sold to various institutions in the country (Laher & Cockcroft, 2013).

The unfortunate demise of the test development capacity at the HSRC in the mid-1990s led to the current shortage of test developers in psychology. This expertise was lost as it was not spread to budding psychologists and thus the knowledge was confined to the spaces in which these experts were found (Foxcroft, 2004). Additionally, most South African test developers are either adapting the international tests used in South Africa (Koch, 2009, 2015) or creating norms for the South African population. Thus, there are only a small number of test developers focused on creating new South African-appropriate tests that could be used within diverse groups (Foxcroft, 2004). In South Africa, it was only in the late 1980s that psychological testing came under scrutiny, thereby emphasising cross-cultural factors (Foxcroft, Roodt, & Abrahams, 2013).

The most prominent factors that were identified as hindering test performance were language and education (Meiring, Van de Vijver, & Rothmann, 2006). The importance of English as the language most frequently used for assessments and the administration of instruments can impact the efficacy of these instruments (McDonald & Van Eeden, 2014). Most Black South Africans have English as their second or third language, yet most psychological instruments are in English, thereby creating a challenge regarding cross-cultural validity (McDonald & Van Eeden, 2014). This challenge has been plaguing psychological testing and test development in South Africa for years, as it remains an immense task to ensure that the tests produced or used are not favouring any individuals or limiting their opportunities unfairly.

There are several important regulations and legislations to be aware of when developing a test. Internationally, the International Test Commission guidelines assist test developers to develop high quality tests. In South Africa, one of these legislations to take note of is the Labour Relations Act 66 of 1995 (Republic of South Africa, 1995), which guides the use of psychometric instruments. The Employment Equity Act (EEA) 55 of 1998 (Republic of South Africa, 1998a; Tomu, 2013; Van de Vijver & Rothmann, 2004) was amended in 2014, as the issues pertaining to test use and development have become a national concern. The Employment Equity Act is taken seriously to protect those completing such assessments. Listed below are the guidelines for the International Test Commission followed by the Employment Equity Act stipulations.

The International Test Commission's (2000) Guidelines for Adapting Educational and Psychological Tests indicated that "test developers/publishers should provide evidence that language use in the directions, rubrics, and items themselves…are appropriate for all cultural and language populations for whom the instrument is intended" (Hambleton, 1994, p. 232).

According to the Employment Equity Amendment Act 55 of 2013 (Republic of South Africa, 2014), section 8:

Psychological testing and other similar assessment are prohibited unless the test or assessment being used a) has been scientifically shown to be valid and reliable, b) can be applied fairly to all employees, c) is not biased against any employee or group and d) has been certified by the Health Professions Council of South Africa established by section 2 of the Health Professions Act, 1974 (Act No 56 of 1974), or any other body which may be authorised by law to certify those tests or assessments. (Section 8 (d) added by section 4 of Act 47 of 2013, Republic of South Africa, 2014; Tomu, 2013)

The recent addition of the fourth requirement in section 8 of the Employment Equity Act was due to a national awareness of test misuse within South Africa. Clause d was however removed in 2017 after a court ruling on the matter. The remaining three clauses however are the requirements that have been the guideline by which tests are used and developed in South Africa to ensure that individuals are not discriminated against unfairly, and thus the developer needs to provide evidence that the test adheres to these requirements. Employers often make use of a cognitive assessment to select potential employees, and this obliges them to ensure that these tests are valid for their intended purposes. In these selections, reading and writing are typically a job requirement and must be assessed, which is often observed in their psychometric assessments. The assessments that are used for employment purposes should be unbiased, implying that it should not unfairly discriminate between individuals. For this reason, the items should be inspected for bias (differential item functioning) as well as the bias in the test (differential test functioning) to ensure that the items and test are not unfairly discriminating, since this could lead to incorrect recommendations being made and individuals losing job opportunities (Tomu, 2013; Van de Vijver & Tanzer, 2004; Van der Pool & Catano, 2008).

Test development is also influenced by situations of power, either enabling or disabling individuals in terms of test performance. This is commonly influenced by the use of cut scores, which is the method employed by test users to decide whether individuals are successful or unsuccessful in a particular test. These cut scores are necessary for decision making and are therefore provided by the test developer to assist test users in this decision-making process. These cut scores are informed by statistical analysis of the data and are normed according to the tested population, to ensure that individuals are afforded a fair opportunity in this decision process (Mahoney & MacSwan, 2005).

The use of cut scores to differentiate high-risk learners from low-risk learners based on language tests can also be problematic, as the researcher must be aware of the standard error of measurement. Therefore, no absolute scores can serve to identify risk groups. Decision making of such a nature carefully considers the validity of test scores and the overall test (Mahoney & MacSwan, 2005). The importance of language testing and its accuracy is evident in the nature of such testing, as one is hesitant to simply classify learners according to their language knowledge. This once again stresses the urgency for good language instruments as it serves to measure vital language and comprehension skills from which inferences about the learner’s language and knowledge must be made.

When considering the injustices inherent in assessing individuals in language tests, the task of critically exploring these issues is indispensable. Critical language testing is grounded in social theory, which attempts to deconstruct visible power relations that assist in creating social or even educational inequalities between individuals or learners in certain contexts. In addition to this, critical language testing theorises that individuals have differentiated access to language, and language is understood as both a resource and practice. Therefore, language testing exemplifies the unequal distribution of linguistic resources, as individuals do not have access to the same resources (Milani, 2007).

The development of measures requires test developers to acknowledge the effect of multiculturalism and multilingualism. This effect cannot be underestimated, as making a measure appropriate and fair for individuals from 11 different languages is a complex task. The exclusive use of English is in itself problematic, as it must be fair for 10 different languages, since this is what will foster a psychologically sound measure. This is the challenge that faces many English assessments in South Africa and the embracing of this dilemma and its solution is what minimises the gap between South Africa and those abroad.

1.3 Overview of the Research Method

This study focuses on the validation of the ECT (Arendse & Maree, 2017), which is a newly developed test for the South African population. The ECT has two test versions, since it was piloted in different years. These two versions are not identical and have to be treated as separate tests; thus, the data cannot be combined. These test versions were, however, compared to assess whether the newer version improved on the first version, indicating that the changes made were effective.

Since this test is still in development, this study is considered exploratory quantitative research, because this was the first time this research was conducted on this instrument. Accordingly, there are no hypotheses stated for the different objectives of the study.

The design and method of this research correspond to psychometric test development and therefore involves the evaluation of the test development and item performance. The methodology used in this study is quantitative due to the range of statistics necessary to gather sufficient evidence of the quality of the test.

The different analyses conducted have assisted in ascertaining whether the ECT has improved from the one version to the next. The various statistical analyses provide more information regarding how the items are functioning and whether the test as a whole is effective. The analyses that were conducted in this study are primarily focused on establishing two characteristics within the test, namely whether the newly developed ECT measures aspects of verbal reasoning, and how well this test measures the verbal reasoning construct.

The aims of this study are to explore the construct validity and reliability of the ECT. This study has five objectives. The first objective is to statistically explore the unidimensionality of items using the Rasch model. The Rasch model is a statistical technique (Rasch, 1960) that allows for an extensive examination of the data to thoroughly explore the items in the test. The Rasch model assess the probability of an individual correctly answering an item, using the parameter of ability (person ability) and a parameter of difficulty (item difficulty). In addition to this, the Rasch model was designed to analyse dichotomous items (which the ECT has) and does so by separately analysing the person ability and the item difficulty. The Rasch model also holds the item discrimination constant across all the items (Hambleton, 1989; Long, 2011; Rasch, 1960; Wright, 1997).

There are many advantages of using the Rasch model instead of classical test theory (CTT). The Rasch model allows one to assess the individual's ability level manifested in the construct (in this study, it is verbal reasoning) based on their responses to test items and on the item properties (Davies, 2003). For this reason, the Rasch model has typically been used in educational settings and more recently, in mental health assessments (Betacourt, Yano, Bolton, & Normand, 2014). A disadvantage of using CTT is that it is sample dependent, whereas the Rasch model is sample independent. Sample independence (in the Rasch model) means that the item difficulty and person ability is calculated separately and is not influenced by the sample used. The Rasch model is based on the theory that an individual's performance on a test is linked to both their ability and the test difficulty (Davies, 2003; Streiner, 2010). This factor is a crucial consideration for this study as convenience sampling was used; the Rasch model allows one to overcome this limitation of sampling by focusing on individuals' ability to complete the test item instead of the sample influencing performance on the test item.

Following this, the construct validity of the ECT is explored. The second objective is therefore to confirm the dimensionality of the test, which involves conducting a confirmatory factor analysis (CFA). By conducting this statistical analysis, the assumption of unidimensionality can be evaluated as it will assist in identifying whether one construct (Verbal reasoning) is being measured by the ECT. This will provide evidence which can support the argument for the construct validity of the ECT. The approach that will be used in this study is structural equation modelling (SEM). This allows one to observe the latent structure of the test as well as examine the individually observed factors that encompass the latent variable (Betacourt et al., 2014).

This method allows one to explore the underlying structure of the test to reduce all the items into groups that represent the constructs of the tests. There are two key types of factor analysis, namely exploratory factor analysis (EFA) and CFA (Martin & Savage-McGlynn, 2013). EFA is the initial evaluation of the data in order to examine the underlying structure. This initial examination is done without prior knowledge of the underlying structures of the test (Martin & Savage-McGlynn, 2013). Since there are two test versions, the ECT test version 1.2 and ECT test version 1.3, EFA was conducted on these two test versions, and it is therefore consequently imperative to conduct a CFA. The second objective, therefore, involves confirming the dimensionality of the test (ECT version 1.3) using CFA. CFA can be used to re-examine the factor structures of the data (Pae & Park, 2006). This objective, therefore, evaluates the dimensions of the ECT version 1.3 which are mostly based on the EFA findings for the ECT version 1.2 through the use of SEM. This will allow the factor structure to be confirmed across the two test versions, since the CFA performed on the ECT version 1.3 will mostly be based on the EFA results of the ECT version 1.2.

The third objective is to provide evidence of construct validity by conducting a multi-trait multi-method (MTMM) analysis. This analysis requires the use of correlations between tests that measure the same construct (which allows construct validity to be confirmed, referred to convergent validity) and tests that measure contradictory constructs and permits discriminant validity to be confirmed (Campbell & Fiske, 1959). This objective is restricted, as a traditional MTMM cannot be conducted as there is information lacking (due to the use of secondary data analysis). A modified MTMM will be conducted with correlations and mono-trait mono-method triangles.

The fourth objective is to explore measurement invariance using differential test functioning (DTF). The DTF refers to statistically evaluating the differences that exist between groups, such as gender and race, in terms of how they performed on the test. DTF is crucial in cross-cultural and cross-linguistic research (Sireci & Berberoglu, 2000). The differences between gender and racial groups are evaluated with the DTF analyses.

The fifth objective is to evaluate the internal consistency of the ECT by conducting a reliability analysis. This reliability analysis is done with the use of Kuder-Richardson Formula 20. This allows one to confirm whether the test is sufficiently reliable and consequently assists in arguing for unidimensionality (but does not imply unidimensionality) within the test. When the items of the test are measuring the same construct, the Kuder-Richardson Formula 20 will have a maximum value (or be closest to 1). This would provide support for unidimensionality and assist in the argument for the construct validity of the ECT.

This study employs Messick's unified theory of validity (Baghaei, 2008; Messick, 1995, 1996; Rambiritch, 2012; Ravand & Firoozi, 2016; Shepard, 1993) as the theoretical framework. The concept of construct validity is unified in this theory and the implications of test use and scores are explored. The consequences of tests usage and score interpretation include the consideration of its relevance, usage, value, and social implications (Shepard, 1993). This paradigm relates to the aim of the study in that it attends to the process of construct validity and considers the social and ethical issues emerging from validating an instrument.

1.4 Overview of Chapters

This study consists of 9 chapters. Chapters One has introduced the study and provided the justification and significance of the research. To contextualise the study, research that is beneficial for understanding practices associated with test development was briefly explored. The implication of test usage and test development was also briefly mentioned. This chapter provided an overview of the research conducted for the study and provides a guideline of the chapters that follow.

Chapter Two concerns research relating to test development since the ECT is a test under development. A short exploration into language history is considered to acknowledge the language aspect of the test. This chapter also includes issues relating to cross-cultural testing, which is an important factor when testing in South Africa.

Chapter Three introduces matters related to cognitive testing, because of its association with verbal reasoning. This chapter investigates the models that have framed intelligence and have influenced the design of many cognitive assessments. The relevance of language in cognitive assessment is explored, and the implications associated with cognitive testing are discussed.

Chapter Four considers research regarding reading and comprehension, which forms part of the language and cognitive aspects of the ECT. The concept of verbal reasoning is introduced as well as studies conducted on verbal ability to further knowledge of the proposed construct of the ECT. A brief background and the initial findings of the ECT are provided, which provides insight into the research conducted for this study.

Chapter Five relates to the theoretical framework that underpins this study; Messick's unified theory of validity. This chapter explores the concept of validity and how it influences test development. The theory behind this framework is explored, specifically the six facets of construct validity. The criticisms levelled at this theoretical framework are also included.

Chapter Six involves the research design and methodology of the study. This chapter is comprised of the primary objectives of the study and the appropriate methodology that was selected to explore these objectives, namely the Rasch model, DTF, MTMM, CFA, and reliability analyses. The ethical considerations relating to this study are also included in this chapter.

Chapter Seven presents of the results of the study. This chapter is comprised of the different outputs relating to the statistical techniques, such as Rasch analyses, DTF, MTMM, CFA, and reliability analyses. This also includes a description of the data and sample. These outputs are presented according to the objectives of the study and the different test versions.

Chapter Eight consists of the discussion of the results of the study. This chapter allows the implications of the various outputs that were presented to be discussed. The discussion is crucial and for this reason, it has been divided into four sections: a comparison between the two test versions on the different analyses conducted, links between the findings and the literature, the use of Messick's six facets of construct validity as the theoretical framework underpinning the study, and a discussion on verbal reasoning as the core construct being measured.

Chapter Nine consists of the recommendations, limitations, and conclusions of the study. This chapter presents the recommendations and limitations made based on the analyses conducted. These insights are important and will be beneficial when planning future research endeavours. Based on all the information examined within this paper regarding the test and theory examined, the summary of these results are presented, followed by the conclusion of the study.

CHAPTER 2: TEST DEVELOPMENT RESEARCH

2.1 Introduction

The ECT is a newly created test and studies relating to test development are significant as they will inform further decisions related to the development of the test. To place this test into context, however, a brief overview of language testing is considered, as this study acknowledges the influence of language on the development of the ECT.

Since the ECT is a locally developed test and should be aligned with international standards, it is necessary to explore international studies on test development. The local literature pertaining to test development is also explored, as one wants to keep abreast with accepted and expected norms. For this reason, legal aspects and social consequences related to test use and development need to be examined.

The consideration of cross-cultural factors is indispensable in test development, especially since the ECT is used in a multicultural context. This requires an exploration into cross-cultural testing that will provide insight into the findings of the analysis. This chapter, therefore, explores research relating to test development, aspects of test development, and the cross-cultural effects of testing.

2.2 The History of Language Testing

There are three stages identified by Spolsky (1975) in the history of testing, but only the first two stages will be explored as they are significant to understanding language testing history (Davies, 2003). The relevance of this exploration lies in the fact that this language history relates to test development history, which is vital in contextualising this study. The first stage identified by Spolsky is “pre-scientific” (Davies, 2003; Giri, 2003; Morrow, 1981), and thus no statistical methods such as validity and reliability measures were performed. This stage sought to emphasise the use of language structures within the test, not realising their lack of analysis would impact the standard of the test. This stage was then followed by the "psychometric-structuralist" stage, as the need to perform these statistical measures became prominent (Davies, 2003; Giri, 2003; Morrow, 1981).

Valette (1967) emphasised the importance of this stage and indicated that examples of good, standardised language tests both developed and administered, were conducted by the Educational Testing Service and College Board. This stage was thus not focused on the language used in the test but more on organising and analysing the test data (Davies, 2003). Moreover culture may influence the items used in the test and the evaluation of the test. These valuable insights into the development of language tests are extremely significant and highlight the prerequisite of good testing practices and the development demand that tests be validated (Davis, 2003). The acknowledgement of items being subject to or influenced by cultural factors is still considered an important variable in test development, this variable thus limits the comparability of tests across cultures or language groups.

Furthermore, the ability of a test to predict the performance of individuals in either work or educational settings is a requirement for the relevant institutes to ensure they are enlisting the correct individuals. This decision-making emphasises the importance of ethical considerations regarding language testing, as these tests are often used for social policy and control. Moreover, the issue of fairness was highlighted as language tests are employed in high-stakes environments. The social status and power of language tests consequently makes them political in nature. Accountability hence lies with the test developer, who needs to be professional and ascertain that the test is suitable for use (Davies, 2003). These issues also apply to aptitude tests, and thus they are pertinent to discussions on test development, test use, and interpretations.

This brief insight into the history of language testing allows one to grasp that the relevance and need to statistically explore the items used in tests, becomes crucial. In addition, the issues around testing individuals became more prevalent as tests became a common method of assessing individuals in educational or institutional settings. These decisions are considered high-stakes and are sources of power, which gives them an important social position. This enforces responsibility and fairness to be practised by test developers and test users.

2.3 Test Development

Psychological measurement endeavours to transform psychological constructs such as intelligence and verbal reasoning into operational constructs that elicit the cognitive abilities of an individual. To do so, however, requires that these psychological constructs be operationalized as measures. This numerical form that approximates measures allows a test to be subjected to various statistical procedures to ensure that the measure is both valid and reliable (Erguven, 2014).

Psychometrics is central to the development of tests as it serves as the method by which tests are evaluated. It provides test developers with methods to statistically examine the usefulness and robustness of their instruments (Martin & Savage-McGlynn, 2013). Various issues need to be considered when developing a test. One of these considerations is test validity, which refers to the test's ability to measure what it was intended to measure. To assess test validity, various statistical methods are used to ensure that the test is performing as it should (Pae & Park, 2006). This validation then implies that it does not prejudice any racial group, gender or language. For this reason, various statistical methods are employed to ensure that the test complies with these standards. Test developers are therefore responsible for performing these different statistical tests to ensure tests that are fair and valid (Pae & Park, 2006; Republic of South Africa, 1998b).

Defining the construct to be measured is essential in the construction of an instrument. Theory informs the construct, and the instrument is created to elicit the construct. It is important to note that the instrument is used to measure a psychological construct, but the instrument could also be measuring other constructs (Sijtsma, 2012). In the exploration of constructs, psychological assessments are often criticised for their use of psychometrics and the techniques used to analyse the constructs. Michell (2000) advocated that psychologists should consider additive conjoint measurement (ACM) (a highly mathematical theory used to quantify traits) to measure psychological constructs, so that these resulting measures are considered more scientific and approach classical measurement. Michell (2008) also views the Rasch model as unequal to the theory of conjoint measurement, because the Rasch model presumes that traits are quantifiable whereas conjoint measurement incorporates ordinal relations which are essential and adequate for the quantification of traits (Michell, 2008; Sijtsma, 2012).

In opposition to this, Sijtsma (2012) argued that psychological measures contain random and systematic errors that violate the conditions by which one can use ACM for analyses and thus prevents ACM from being an appropriate method of inquiry. As a result, Sijtsma (2012) promoted the use of item response theory (IRT) (a statistical analysis used to examine the items of a test) as it can assist with the scientific inquiry of psychological measures. He argued that IRT had the same goals as ACM, but unlike ACM, IRT (including the Rasch model) made provision for systematic and random errors within these measures (Borsboom & Mellenbergh, 2004).

This study identifies with the argument for the use of IRT (Sijtsma, 2012) and has as such used the Rasch model to further the development of the ECT. Linking with the argument made by Sijtsma (2012), the use of Rasch analysis as a useful and comprehensive method for examining tests is advocated. Rasch analysis is acknowledged for its assistance in the development of instruments, especially those pertaining to ability and achievement. The Rasch analysis method allows the items and test information to be examined as well as the model fit, which assesses the reliability and validity of the test (Rasch, 1960; Els & Andries, 2011). Rasch analysis is, therefore, an effective method of analysis when evaluating the development of tests.

The Rasch model can be regarded as an example of conjoint measurement, in that it is a practical method of applying this theory to empirical data. Moreover, the Rasch model complies with a necessary requirement of additivity. The Rasch model can be considered as additive, as the person ability and the item difficulty, which are considered to be two independent variables, can be measured on the same scale with equal intervals. Although the Rasch model bears great similarity with conjoint measurement, it however differs from conjoint measurement in that it includes probabilities which allow the model to integrate measurement errors since empirical data is subject to error (Acton, 2003; Baghaei & Amrahi, 2011; Smith, Wakely, De Kruif, & Swartz, 2003; Wright, 1997).

Within psychology, tests such as the sentence completion test are used for clinical purposes. This test is used for people between ages 8 and 25 and was developed on the theory of ego development (Els & Andries, 2011). From this analysis, the chi-square fit statistics, which includes the infit and outfit mean squares were observed to be in an acceptable range, indicating that the data fitted the model. The person-fit information allowed the researchers to argue that items were not biased and that there were only a small proportion of unrelated responses. The Rasch person reliability was very high, indicating that the test could be trusted to provide an indication of the candidate’s ability. In addition to this, Rasch factor analysis, which involves exploring the shared variances in the residuals, was performed to assess the unidimensionality of the instrument. Lastly, the eigenvalues were considered (< 2) and based on these findings, construct validity was established for the test. This study demonstrated valuable insights into the reliability and validity of a test when performing a Rasch analysis (Els & Andries, 2011).

2.4 International Test Development

In the United States of America, the use of the Standards (American Educational Research Association) has guided the development and use of psychometric instruments. The Standards is not a legal document, but rather a guideline for professionals interested in producing quality tests. The American Standards published a section regarding how test scores should be interpreted and indicated the five elements that should be present or examined, namely “test content, response processes, internal structure, relation to other variables, and consequences of testing” (Sireci, 2007, p. 478).

According to the American Standards, the evidence of these elements should be reported for a test to be considered valid. The bodies of evidence required by the Standards for the five elements are comprised of test content information, which can be provided as a form of job analysis, and the use of subject-matter experts, which informs the content of the test items. In education, the concept of alignment is used, and this is also considered a source of evidence (Sireci, 2007; Sireci & Parker, 2006).

The evidence required for the response processes includes observations on response patterns and interviews with individuals on their response patterns, which allows one to identify a link between the construct and the response pattern of the individual. The evidence required for assessing the internal structure of the test is examined using exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) to assess the structure of the data. The evidence needed for confirming relations to other variables, which is also referred to as criterion-related validity, consists of correlations with relevant criteria and multi-trait multi-method (MTMM) analysis, which informs one of the relations to other similar tests and predicts performance. Consequences of testing involve assessing the issues that may emerge from test administration, development, use, and interpretation. The consequences include adverse impact and high failure rates (Sireci & Parker, 2006). The above-mentioned sources of evidence such as CFA and MTMM correspond to the evidence required by this study. Furthermore, the theoretical framework of Messick's unified validity theory, which was used in this study relates to Messick's requirement of assessing the consequences associated with test usage and interpretations made. These standards, therefore, outline requirements made by Messick's unified theory of construct validity.

The American education system has devised a process of alignment to ensure that the same construct is measured in all state examinations, which individuals must have completed at school level. By ensuring this, policies can be influenced more profoundly, and all those involved in the education process have a clear goal. The aim of such alignment is to increase the performance of individuals on tests thereby inferring that they grasp the relevant content and skills. This alignment must be demonstrated as an agreement or similarity between the expectations of the state examinations and the requirements of the Standards (Resmick, 2003).

There are several means by which this alignment is evaluated, such as exploring whether the content corresponds to the Standards, whether there is a range of knowledge being assessed, if it is cognitively appropriate for the group being tested, and that no irrelevant information is assessed in the examinations. Furthermore, the educational standards of the states in America are influenced by the "No Child Left Behind" Act, which encourages them to create tests that are aligned with their state's standards for learners. The elements with which the examinations and standards must be aligned is, however, not easy and still requires work. It is, nonetheless, a noteworthy effort that American (United States of America) education is endeavouring to ensure that all their 50 different states produce learners that are aligned with what they should know to continue studying after school (Resmick, 2003).

In the United States of America, there are a few laws that are used to challenge the use of a test, namely the Civil Rights Act of 1964 and 1999 (Title VI), the Civil Rights Act of 1964 and 1999 (Title VII), the Equal Protection Clause of the fourteenth Amendment (United States Constitutional Amendment fourteenth), and the One Process Clause of the fourteenth Amendment (United States Constitutional Amendment fourteenth). These laws are commonly used for issues relating to adverse impact, in that majority groups are advantaged over minority groups (Sireci & Parker, 2006).

Test discrimination across cultural groups in the USA is defined according to three regulatory bodies. The first is the United States Equal Employment Opportunity Commission (EEOC) that uses the “four-fifths rule” in the selection of individuals - this implies that minority groups have a lower pass rate and disparate impact is observed. The second is the Supreme Courts standard deviation analysis, which indicates that if there are three or more standard deviations between two different groups, then there is a disparate impact present. The third is the “Shoben Formula”, which involves observing a statistically significant difference between majority and minority groups scores which would indicate that a disparate impact is present (Sireci & Parker, 2006).

There have been some documented cases in which psychometric tests have been used in court proceedings and where their development and use had to be defended. Sireci & Parker (2006) examined four court cases that involved tests for employment and educational purposes. The court procedures unfolded as the plaintiffs accused the use of the test to be discriminatory towards minority groups. The plaintiff proved the accusation and the defendants were approached to prove that the test was not discriminating based on job requirements and it consisted of relevant job content. If the defendants proved that the test was not discriminating unfairly, the plaintiffs could appeal and suggest the use of another test for assessment purposes that do not discriminate unfairly and that also assess the job requirements (Sireci & Parker, 2006).

The reasons for the plaintiffs taking the different tests to court commonly revolved around group differences perceived as disparate impact and discrimination. The evidence provided by all of the court cases was based on content validity and consequential validity. Three out of the four court cases included external correlations (showing predicted performance) and only one of the four court cases supplied evidence of the internal structure of the test. This evidence was sufficient for the courts to make a decision, and the test evidence provided in each court case was deemed as legally compliant. It is interesting to note that all cases did not use multiple sources of evidence as prescribed by the psychometric guidelines in the American Standards guiding test validation. Although this may be perceived as lenient in terms of the court’s requirements for legal cases regarding test validation, individuals involved in psychometric test validation should be well versed with procedures and guidelines as they may be requested to be an expert witness in court cases (Sireci & Parker, 2006).

One legal case of interest was the “Golden Rule Case”, which showed that Black candidates were disadvantaged in comparison to their White counterparts regarding the items of the test. This led to differential item functioning (DIF) analysis as a means of exploring group differences on items. The importance of reviewing court cases and legal requirements lie in the fact that any test can potentially be accused of having a disparate impact and the evidence acquired by the developer should be able to withstand legal scrutiny and be aligned with psychometric guidelines (Sireci & Parker, 2006).

The Canadian Forces need to adhere to their Canadian Employment Equity Act, which requires them to employ individuals from different racial groups. This was, however, causing a problem, as many of the non-White applicants were not meeting the minimum requirements on the cognitive assessment in their selection procedures. This caused officials to investigate the issue, as they were not sure if it was due to biased instruments or a lack of ability in the individuals. Multiple cognitive measures were evaluated to ensure they were assessing the same dimensions across groups. The study found that the verbal cognitive ability assessments allowed for larger than one standard deviation between Native North American and Whites, while the non-verbal assessments also displayed differences (< 1 standard deviation) (Van der Pool & Catano, 2008). On the non-verbal measures, they concluded that the individuals had the same cognitive ability. They found only a few biased items in some of the cognitive tests, which allowed them to speculate on the inferences made for the selections using these tests. The adverse impact of the verbal cognitive tests was assessed and all had an adverse impact, implying that the Native North American applicants that were successful were less than four fifths of the successful White applicants in the selection. The conclusion drawn from the study was that the differences between Native North American and White applicants existed due to language, culture, education, and ability. Some of the verbal measures, when controlling for education and language, unbiased and differences were then also related to ability. Since the verbal measures had an adverse impact on selection decisions, non-verbal tests that measure the same construct were suggested as an alternative as these may be a less biased instrument for selection decisions (Van der Pool & Catano, 2008).

In an article by Ardila, Ostrosky-Solis, and Bemal (2006), four criteria were identified and proposed for establishing valid and reliable cognitive measures. The first criterion was defined as normative, which involves the population being tested. This includes the consideration of an individual's culture, age, education, language, and so forth as variables that affect performance on the test (Ardila et al., 2006). This criterion is usually identified by the demographic information obtained by the sample being studied. These factors can, however, be explored in more depth to ensure the assessment considers the sample when validating it. The second criterion, known as the clinical criterion, requires the relevant possible brain pathology which could impact the test to be identified. There should also be an awareness of cognitive disabilities that could inhibit performance (Ardila et al., 2006). This is often not considered when some cognitive tests are developed due to the nature of the construct being examined.

The third criterion was called the experimental criterion, and it similarly involves exploration of the brain, and links the test performance to stimulated areas in the brain. This interesting aspect is commonly demonstrated by using brain imaging and identifying the parts of the brain that are activated when individuals are completing the assessment.

The fourth and final criterion was labelled the psychometric criterion, and emphasises the need to know how the test relates to other cognitive tests. This relates to the practice of validation, which requires the researcher to evaluate the quality of the assessment measures. This criterion also emphasises the importance of establishing valid and reliable measures. Although most cognitive tests do not adhere to the above criteria, it is a proposed process that test developers of cognitive assessments can follow to assist in establishing a valid assessment measure (Ardila et al., 2006).

Ardila et al. (2006) used the Semantic Verbal Fluency (ANIMALS) Test to demonstrate the effectiveness of the four proposed criteria. The information obtained on these criteria provided a comprehensive picture of the test and how it was performing. The effectiveness of exploring the different aspects of a test can allow for an inclusive interpretation of the test. Additionally, in terms of cognitive test construction, it was suggested that a multiple-choice format be used as a method of answering items, because it reduces methodological issues such as affective influences and memory (Langdon, Rosenblatt, & Mellanby, 1998).

In the USA, several studies have been conducted to explore the predictive validity of cognitive ability tests. They found that these tests were valid for the USA population and were able to predict job performance in various environments and occupations. These findings can, however, only be generalised to the USA population and have not considered the demographics of other populations that utilised these cognitive tests. Many tests, such as the differential aptitude test (DAT), are used in the United Kingdom and several American-developed practices have been utilised in the UK. The UK utilises more cognitive tests for selection decisions than in the USA. The emphasis placed on test evaluation in terms of test construction and validation in America encouraged the UK to evaluate their test usage and validate their instruments (Bertua, Anderson, & Salgado, 2005).

The selection practices in the UK favour the use of specific cognitive tests such as verbal and numerical tests. However, the USA predictive studies found that specific cognitive abilities such as verbal ability had less predictive power than general cognitive ability in determining occupational performance and success (Bertua et al., 2005).

These issues emerging from international research on test development, legislation regarding test usage and development, as well as the use of cognitive tests in selections are critical to advance understanding on these matters. This engagement allows one to consider whether South Africa is obtaining similar findings or whether the American and British findings contradict the test development practices in South Africa. Nonetheless, the issues raised allow one to consider their importance and the dire consequences associated with unfair test usage.

2.5 Cross-Cultural Testing

A three-dimensional view of culture endorsed by Li (2003) consists of the following facets: resource, process, and developmental relevancy. The resource facet of culture considers that certain elements, such as knowledge and values, grow over time, are closely influenced by society, and have an impact on individuals. This resource facet was further developed by Willis and Schaie (2006) in that it included aspects such as education and occupational achievements attained by the individual. As a result, this facet corresponds to the stage of early adulthood whereby achievements are dependent on factors such as learning, cultural information, and skills. The subsequent facet denotes the process of culture that consists of behaviour attributable to everyday life, which involves individuals engaging in routine behaviour. This behaviour is usually influenced by societal aspects and the individual's immediate environment (Li, 2003; Schaie, 2006, 2008).

The last facet is labelled the developmental relevancy of culture because it encompasses both the processes and resources with which the individual is engaging. This facet corresponds to the developmental stage of the individual, which essentially positions them in terms of their age-related progress (Li, 2003; Schaie, 2006, 2008). This view of culture allows one to grasp the complex nature in which culture is inherently part of people's lives and this shapes them both cognitively and emotionally. Since culture is part of an individual's development, it becomes indispensable to include this aspect when considering test development (Malda, Van de Vijver & Temane, 2010). The discussion of issues relating to culture and its impact on testing individuals is massive, thus it is necessary to consider this aspect in test development.

When considering cross-cultural testing there are several elements that need to be considered. These elements involve the instruments utilised, the methodology used, and the interpretations made for diverse populations (He & Van de Vijver, 2012).

In cross-cultural testing, there are three alternatives available concerning the choice of instrument to be utilised, such as adoption, adaptation, and assembly. The use of instruments through adoption refers to instruments that have been used in one cultural group and are then applied to another cultural group. For example, American or European developed tests are used in South Africa for educational or occupational reasons. This method of adoption is usually regarded as a simple process and allows greater opportunities to compare scores. It is, however, limited in its ability to make substantial comparisons, as the construct and items being measured across groups need be established as equivalent for all cultural groups using the test (He & Van de Vijver, 2012).

The second available option is adaptation. This refers to the test being transformed into another version by changing the language, and in certain instances, allowing some items to be transformed more profoundly to retain the same construct being measured. This often occurs in cases where direct adaptation might create problems linguistically and psychologically. In the past, there was more focus on translating items and tests to be linguistically sound in another language, and consequently the psychometric properties was often overlooked. Presently, both linguistic and psychometric aspects are considered when tests are adapted from one language or culture to another. The third alternative is assembly. This refers to the construction of a novel instrument, which is usually considered when current instruments are not adequately measuring the intended construct. This option will allow an instrument to be directly applicable to the target group being measured, but might not be applicable in settings outside of its construction (He & Van de Vijver, 2012). The construction of the ECT can be referred to as application of the assembly method.

Selecting one of these three methods is guided by the reasons for which one wants to use the instrument. When the choice is based on increasing opportunities to perform statistical comparisons, then adoption would be the preferred method. In cases where the emphasis is on sufficiently measuring a particular construct in a different culture or language, then one would consider either adapting or assembly methods. Moreover, statistical techniques such as IRT and SEM are used to deal with cross-cultural testing. When using such techniques, if it is found that there are more items in the instrument that are unique to the culture than those that are not; comparing the items across cultures will be limited, if not inaccurate. Thus, when using instruments across cultures or for cross-cultural testing, the ability to make comparisons across cultures may not increase the ability of the test to be locally valid (He & Van de Vijver, 2012). The two methods that are promoted for examining cross-cultural tests are used in this study and will therefore provide substantial information about the construct and items in the ECT.

There are three forms of bias, namely construct bias (invalidity within the construct being measured), method bias (an invalid instrument), and item bias (invalid items). Construct bias refers to an incomparability between the construct being measured across cultures. This is observed when the construct is not fully measured in both cultures, as some aspects are missing. This implies that the way in which the construct is defined in both cultures is limited, because they are only slightly compatible. Those embarking on cross-cultural testing are therefore required to adequately define the construct being measured. Consequently, when defining the construct, all associated meanings must be considered as well as all aspects that possibly constitute the construct. This also refers to how different cultures perceive the same construct, as this may be where the difficulty lies. When considering such aspects, the limitations associated with the construct across cultures must be recognised (He & Van de Vijver, 2012; Van de Vijver & Tanzer, 2004). Since construct validity is one of the aims of the present study, the recognition of possible construct bias must be established. This allows one to be certain that the construct being measured for the ECT is unbiased and cross-culturally appropriate.

Method bias refers to problematic factors that occur in the sampling, the instrument itself, or administration procedures. The first is sample bias, which occurs when the samples differ significantly from each other and consequently limits the ability to compare the samples. The samples can differ with regard to several issues such as education, urban and rural (location), age, or religion, depending on the construct being assessed. This implies that when making comparisons across cultures, the samples should have similar, if not the same characteristics. It is worth noting that there are differences between the education and language of developed and developing countries, which make cross-cultural comparisons difficult. This would imply that one cannot simply compare different populations with each other as the language used and their education system may differ. This would then cause a biased comparison. Sampling methods may also affect the comparability of samples; for example, convenience samples limit one's ability to generalise the findings of the sample to the population. Thus, data collection should be done in such a method as to maximise sampling and generalisability. When constructs such as intelligence or cognitive ability are assessed, the sample being used should be controlled for characteristics such as education, because these might become unintended variables when statistically exploring the test across cultures (He & Van de Vijver, 2012; Van de Vijver & Tanzer, 2004).

Instrument bias refers to issues within the structure of the instrument that prevent individuals from answering easily. Educational and cognitive testing may have problems regarding whether the target groups are familiar with the elements being tested. For example, different cultures are familiar with different aspects of language and may struggle with the language and cognitive content of tests, as it may not be familiar to them. The unfamiliarity experienced by individuals can also be linked to their unfamiliarity with different response procedures used in the test (assessment). The response procedures used in assessments refer to the different answering options. Studies have also found that background factors affect an individual's performance on cognitive assessments (He & Van de Vijver, 2012; Van de Vijver & Tanzer, 2004). Another study (Malda, Van de Vijver, & Temane, 2010) found that when two test versions, Afrikaans and Setswana versions, was used in South Africa, the respective cultures achieved better scores when assessed in their respective test version language. The suggested approach to addressing issues stemming from familiarity is that tests be adapted within the target cultures or languages (He & Van de Vijver, 2012; Malda et al., 2010).

Bias associated with response styles refers to problems that occur when individuals answer in a particular way. For example, consistently choosing yes responses opposed to no responses (this is termed as acquiescence in personality assessment) is what causes the bias as the answering pattern takes preference over the construct being measured. A method of assessing whether response styles are influencing the performance of individuals in the assessment is done by the use of correlations. Correlations between the corrected scores of within-individual and within-cultural standardised raw scores are conducted, and depending on the magnitude of the correlation, evidence of possible response style bias may be detected. These correlations are, however, not a definite affirmation of the bias of response styles and therefore need to be conducted with caution. These differences could be indicative of real differences and not response styles, which could be due to cross-cultural differences. The response styles should then not be changed until all aspects are explored, so that the real differences across cultures are not influenced by the removal of response styles (He & Van de Vijver, 2012; Van de Vijver & Tanzer, 2004).

Another aspect of bias is administration bias. This results from several circumstances such as the method in which the data were collected; unclear, or confusing instructions during administration; the way in which the candidates and the administrator function (halo effect); and difficulties associated with language, communication, or both. The method, in which the test is administrated, such as pen and paper or computer-based programs, could affect how individuals answer, depending on their level of familiarity and need to give socially acceptable answers. Method bias can affect the scores of individuals on assessments, and acknowledging issues in the measurement when analysing the data could explain possible score differences instead of assuming cross-cultural differences (He & Van de Vijver, 2012; Van de Vijver & Tanzer, 2004).

The last form of bias is item bias, which refers to an item having a dissimilar meaning across different cultures. This is observed when an item receives different responses from individuals who have the same ability but are not from the same cultural group. This enforces the idea that the item is biasing certain individuals or cultural groups. Issues relating to biased items are usually due to the following: translating or adapting items, differences in the content of the item across cultures, items that have several meanings in another culture, or some items that do not exist in another culture or language (He & Van de Vijver, 2012; Van de Vijver & Tanzer, 2004). The last aspect mentioned of the absence of some constructs across cultures is what motivates the construction of the ECT, as there are many English concepts that do not exist in African languages (Koch, 2015). This links to problems experienced in verbal analogy assessments in South Africa (Koch, 2015).

Equivalence is an important concept when adapting tests into different language versions. Equivalence is regarded as a hierarchical process, which is comprised of construct, measurement unit, and full-score equivalence. Construct equivalence refers to the same construct being measured across the cultures. This is an initial standard by which cross-cultural comparisons are made, because it implies that the same construct is present across cultures. If the same construct is not observed across cultures, then they cannot be compared and this suggests that the construct must be re-visited and explored in more depth. The second level of equivalence is measurement unit equivalence (metric equivalence), which refers to assessments that have the same measurement level (such as interval or ratio) but have dissimilar origins. This suggests that when metric equivalence has been established, the scores may be compared within cultures. The last level of equivalence is full-score equivalence (scalar equivalence), which is the top level. This refers to assessments that are the same in terms of where it originated from and the measurement units used. This implies that cross-cultural comparisons are possible as scores are equally comparable, meaning scores are unbiased. Fundamentally, construct bias affects construct equivalence, while method and item bias affects both measurement unit and full-score equivalence (He & Van de Vijver, 2012; Van de Vijver & Tanzer, 2004).

Furthermore, to reduce cross-cultural bias, instruments should be examined vigorously using psychometric methodologies and an individual familiar with the culture should interpret the data being examined. Thus, the effects of culture should always be considered when interpreting the results of an instrument, thereby limiting individuals from different cultures from being disadvantaged (Tseng, 2001; Van de Vijver, & Rothmann, 2004).

The influence of demographic factors such as age, gender, race, occupation, education, urban or rural origins, and hometown can affect how individuals perform on measures of intelligence (Jensen, 1974; Van der Pool & Catano, 2008). A formula known as the Barona demographic formula (Barona, Reynolds, & Chastain, 1984) was developed to measure the relationship between cognitive ability and demographic influences. It is a regression model and is dependent on known information (Van der Pool & Catano, 2008).

The research related to the use of the formula was found to be reliable for ill and healthy individuals. The formula is presented below to demonstrate how these different factors were taken into consideration to generate an IQ score.

IQ = 54.96 + 0.47 (age) + 1.76 (gender) + 4.71 (race) + 5.02 (education) + 1.89 (occupation) + 0.59 (region).

The rural and urban divide regarding performance on cognitive assessments has been well researched, because it was previously found that individuals from rural areas performed better on non-verbal and spatial tests compared to verbal tests, giving urban individuals an advantage in verbal assessments. These changes are, however, not as pronounced as in the past and researchers often overlook this as this demographic variable has little influence on cognitive performance (Van der Pool & Catano, 2008).

In terms of gender differences in cognitive abilities, it was found that females had higher scores for verbal ability (reasoning) and other verbal-related cognitive assessments such as word memory, anagrams, writing, general, and mixed verbal ability assessments. Males scored better on verbal analogies and spatial relations tasks. There were, however, very small differences between males and females in terms of general cognition or IQ (Griskevica & Rascevska, 2009; Strand, Deary, & Smith, 2006).

This consequently dismisses the idea that either gender is more intelligent. A longitudinal study (Camarata & Woodcock, 2006) demonstrated gender differences in that females performed better under time pressure while males scored higher in verbal tests assessing analogies, antonyms and synonyms, and identification (Griskevica & Rascevska, 2009). This finding is interesting as the ECT contains antonyms and synonyms, and it would be interesting to determine if the same gender performance is observed in the ECT scores.

Based on historical discrimination, it is expected that group differences exist between African Americans and White Americans. A study by Whitefield, Allaire, Gamaldo, and Bichsel (2010) indicated that the verbal meaning (vocabulary) and inductive reasoning tests were invariant across ages taking into account the educational opportunities for African Americans, with emphasis on the quality and not the quantity of their performance. This study suggests that the influence of education cannot be ignored when comparing cognitive abilities across age groups of African Americans (especially older individuals) (Whitefield et al., 2010).

A study by Flanagan and Ortiz (2001) explored the performance of aboriginal children from Canada on the WISC (Wechsler Intelligence Scale for Children); their results revealed that these children performed very poorly on this test. Researchers found that the verbal scale loaded highly on linguistics and culture, which was not familiar to the aboriginal children. These, among other factors, explained their poor performance on the test. The unfortunate consequence of such tests being used on individuals such as the aboriginals is that their intelligence is inaccurately measured, which results in them being disadvantaged (Flanagen & Ortiz, 2001).

The reliability of psychometric instruments is important when evaluating the performance of non-English individuals on English measures, but establishing the validity of the instrument takes precedence (Sotelo-dynega, Ortiz, Flanagan, & Chaplin, 2013). Research done on the performance of bilinguals and monolinguals on the Peabody Picture Vocabulary Test and Ravens Colored Progressive Matrices and the dimensions that emerged from these tests emphasized the importance of validity. The results of the study indicated that the verbal factor emerging from the bilingual group could possibly be biased. This finding suggests that other factors within the bilingual's culture need to be considered when exploring their performance on assessment measurements (Sotelo-dynega et al., 2013).

An article by Budd (1998) explored test usage in the UK, specifically that of a personality measure. His findings indicated that British test publishers were selling American personality tests without considering the cross-cultural effect these tests could have on the British people. These test publishers, however, sought to make these personality tests cross-culturally valid by "anglicising" them. His criticism on anglicising (which refers to changing the American words used in the test to British words) lies in the fact that one cannot be sure that the anglicised forms of the test are equivalent to the original American form (Budd, 1998). Additionally, there has been no evidence provided to disprove his criticism. Since multinational companies use these personality tests, the emphasis on cross-cultural validity and research in test usage is important. The fact that individuals taking the tests are different in terms of their cultural backgrounds will affect how they respond to items in the test is a concern that should be considered very seriously (Budd, 1998).

Cross-cultural studies identified that African Americans performed poorer on cognitive assessments than White Americans. This was explained by historical inequalities and low education achievement, which has led to their limited educational advancement and has restricted them in their advancement in employment opportunities (Kennedy, Allaire, Gamaldo, & Whitfield, 2012).

Interesting findings were discovered in Kennedy et al.'s (2012) study on the intellectual control beliefs in older White and African Americans. Firstly, older African Americans performed poorer than the White Americans in all the cognitive tests and they had low internal and high external control beliefs. These control beliefs are associated with low intellectual control. Intellectual control beliefs were then assessed with other African Americans and the same findings were made, regardless of age, gender, or educational level (Kennedy et al., 2012). This is a noteworthy aspect to consider when assessing individuals such as non-Whites in South Africa. Their performance on cognitive assessments could be influenced by their control beliefs, which could impact their ability to perform better than their White counterparts.

Research by Abrahams and Mauer (1999) on the South African version of the 16-Personality Factor (16PF) Questionnaire indicated that the Black (African) group differed from the White group in terms of their response patterns and internal consistency of primary factors. They attributed this to language proficiency and cultural factors. Another study was then conducted which also focused on the South African version of the 16PF in terms of its cross-cultural use, with a specific focus on the vocabulary used (McDonald & Van Eeden, 2014). The results from this study on Black (African) and White university students indicated that generally they performed similarly, yet differences were observed for the Black group in their level of vocabulary and their overall lower mean score (McDonald & Van Eeden, 2014; Van de Vijver, & Rothmann, 2004). This study thus emphasised the effect of language in psychological measures such as the South African 16PF on the African population.

The study on the Learning Potential Computer Adaptive Tests, (LPCAT), a non-verbal test developed in South Africa showed evidence of cultural bias (De Beer, 2004). The results obtained in this study illustrated that tests in either verbal or non-verbal form can contain effects of culture and influences of education. Consequently, bias is explored by examining equivalence, which is when multiple versions of a test exist and DIF analysis is conducted to assess how the items function across different language groups (Van de Vijver & Tanzer, 2004; Van de Vijver & Rothmann, 2004). Interestingly, another study on the LPCAT indicated that the subtest, verbal reasoning, was the best at predicting academic performance (De Beer, 2011).

A project known as the ABLE project (Additive Bilingual Education) was developed with the purpose of implementing the model of additive bilingual education and sought to empower learners’ primary language, encourage literacy of learners in two languages, and ensure learners become academically strong and competent in two languages. These objectives were implemented in a rural area of the Eastern Cape, South Africa, where isiXhosa was identified as the primary language of the community and was the language of instruction of the school, with English as an additional language (Arendse, 2010; Koch, 2009).

The Woodcock Munoz Language Survey (WMLS) is an instrument that measures the development of academic language proficiency in an individual's primary and second languages, and is used extensively in the USA to evaluate children's Additive Bilingual Education programmes and language proficiency in English and Spanish (Woodcock & Muñoz-Sandoval, 2005). The ABLE project members intentionally selected the WMLS to evaluate the language development of learners because it allowed them to assess the language outcomes of the project, thereby measuring the selected participants' performance in language as well as evaluating the effectiveness of the additive bilingual programme (Arendse, 2010).

The WMLS was therefore adapted into South African English and isiXhosa to assess the academic language proficiency in English and isiXhosa of the learners in the project. Research on the equivalence of the two language versions of the test was the next step in the process of translating and adapting the test into isiXhosa. The study on the verbal analogies subscale was important in the general adaptation of tests into the indigenous SA languages because it was initially found to be a problematic test and a completely new subscale had to be developed (Koch, 2009). There was a complete change from the direct translation method to the re-writing and adaptation of the scale into culturally appropriate language, but still tapping into the same underlying psycholinguistic construct. As a result, a number of items in the isiXhosa version differed completely from the original version, yet it still produced very promising results (Arendse, 2010; Koch, 2009).

An exploratory factor analysis was conducted and the results obtained from the factor analysis in the verbal analogies subscale indicated that only factor one could be regarded as structurally equivalent, while the second factor was measuring different constructs across the two language versions. The implication of the first factor was that the items could be used for comparison across the two groups. The second factor presented problematic loadings in the isiXhosa version, and was not equivalent across the versions, supporting the findings of the previous research (Koch, 2009) of possible differences in the weightings across the two language versions on one of the dimensions. This implied that the items observed for the second factor were not comparable across groups (Arendse, 2010). Moreover, the second factor displayed construct bias, due to the inconsistency in the overlapping of constructs across the two language groups (Meiring, Van de Vijver, Rothmann, & Barrick, 2005).

The ABLE project (Koch, 2015) demonstrates the difficulties associated with adapting instruments across languages in South Africa, as well as the issues that arise when attempting to measure constructs formulated in English in African languages. The study also highlights the difficulties mentioned by He and Van de Vijver (2012) when adapting instruments from one context to another. This therefore highlights the need for locally produced measures, such as the assembly method that might resolve such testing difficulties.

Cross-cultural testing is not an issue specific to South Africa, as the previous international studies indicated similar issues with indigenous individuals. As a result, one cannot underestimate the influence of how an instrument is developed, used, or administered, bearing in mind that the barriers surrounding test development and use are international and inherent. This requires test developers to be more vigilant in how these instruments are being used both locally and internationally, as well as the populations on which they are being used.

2.6 Conclusion

This chapter sought to contextualise the study in terms of the history of language testing. It allows one to understand how the measurement of language began and the issues which required the use of psychometrics to evaluate the effectiveness of language.

The importance of understanding test development and the issues related to developing tests were indicated in this chapter. These issues needed to be investigated to contextualise the ECT as a test in development. To supplement this, the relevant literature relating to international and national validation studies was examined to establish the lessons learnt and the possible implications.

The research relating to cross-cultural factors emphasised the numerous factors that could potentially bias instruments and the importance of decisions relating to the use of instruments in cross-cultural testing environments. The awareness of these factors is imperative when assessing instruments in multicultural contexts and informs one of possible reasons for discrepancies in the data analysis.

The various case studies pertaining to test development and cross-cultural testing provide an examination of relevant examples and situations that must be considered when developing a test. These considerations include limiting the possible extraneous variables that might affect one’s study without careful thought.

CHAPTER 3: COGNITIVE TESTING

3.1 Introduction

When evaluating the issues presented in the previous chapter on test development, there is great emphasis on the construct, as it informs the items of the test. In this study, the construct of verbal reasoning is of particular interest and requires one to gain an understanding of this cognitive skill. Since verbal reasoning is a cognitive construct, it is imperative to explore the nature of cognitive testing as well as the theories underpinning the field.

Cognitive tests involve thinking, which is characterised by reasoning, memory, verbal, and mathematical aspects. Tests attempt to measure these constructs by using items that tap into these thinking systems. Tests are therefore comprised of items which operationalize the construct. Assessing whether the test, which is comprised of items, approximates measurement is crucial. Thus the two processes of ‘understanding the construct’ and ‘measuring the construct’ occurs in either the respective order or they happen concurrently.

When testing for cognitive abilities, the test can either have different measures of intelligence or be focused on specific forms of intelligence. Cognitive tests assessing general cognitive ability have been judged to infer learning potential as well as to predict job performance (Hunter, 1986; Kvist & Gustafsson, 2007; Lohman & Lakin, 2009; Sternberg, 1986). Cognitive assessment focusing on verbal intelligence includes constructs such as reading comprehension (Kendeou, Van den Broek, Helder & Karlsson, 2014) and verbal analogies. The most commonly used item formats involve multiple-choice, true or false, and sentence completion item questions. These are similar to the item formats observed in the ECT.

This chapter explores the theories of intelligence under which the construct of verbal reasoning is found. These theories are valuable for one to understand how verbal reasoning works and the way in which it is executed in assessments. This chapter will explore the psychological theories of cognitive development, models of intelligence, the influence of the intelligence models, and language and cognition. These factors relating to cognitive testing are vital and extend one's understanding of how these cognitive structures interact.

3.2 Psychological Theories of Cognition

The utility of the brain needs to be acknowledged at the core of understanding how individuals learn or use language. The process of learning occurs through the activity of neurons and synapses in the brain. The brain is constantly adapting and stimulates simple and complex thought. The brain continues to change throughout one's lifespan (Ormrod, 2008). This insight into how the brain functions are what aroused curiosity about how and why humans reason the way they do. This curiosity resides squarely in the domain of psychology as the theorising of how and why humans do what they do is what prompts psychologists to engage with these philosophical questions.

Cognitive tests are commonly informed by the models of intelligence, but the impact of psychological theories on cognitive assessment is equally important as it forms the lens by which this psychological construct can be understood. Psychological theories try to explain human development from birth to death as well as brain functioning within this period.

Various aspects of functioning must be explored to fully understand an individual's functioning in tests. Within developmental psychology, aspects such as biological, emotional, physical, cognitive, social, and personality are explored to provide a comprehensive picture of an individual's functioning. These aspects change throughout an individual's lifetime and hence psychologists focus on the different life stages from birth to death. These developmental frameworks allow psychologists to intervene and address individuals more accurately by considering all these various aspects in terms of their age groups (Blake & Pope, 2008).

Various programs, such as No Child Left Behind in America, were created within a developmental framework, which attempted to increase the number of children achieving at school. It had a specific focus on their cognitive development. In this study, the element of development under consideration is intellectual functioning, which falls within cognitive psychology. This means considering mental processing, thinking, perception, memory, and learning abilities, which fundamentally involve activities of acquiring, processing, and storing information (Blake & Pope, 2008).

The two theorists who were instrumental in from a psychological perspective capturing human cognitive development were Jean Piaget and Lev Vygotsky. A brief investigation into these theorists' approaches is crucial to understanding how cognitive development was understood and how it influenced the development of cognitive assessment. These theories present one with the insight that is required to understand performance on assessments of ability and intelligence which leads to inferences about assessment and testing.

3.2.1 Jean Piaget's Stages of Cognitive Development

Jean Piaget (Piaget & Cook, 1977) worked in Alfred Binet's laboratory, which is where his interest in cognitive functioning in children arose. His interest was, however, directed at understanding why individuals, specifically children answer questions incorrectly and whether these incorrect answers were related to a lack of knowledge. He believed that their incorrect responses were not random and hence the subsequent stages that individuals progress through are related to acquiring information (Blake & Pope, 2008; Piaget & Cook, 1977).

As a result of his investigation, Piaget developed four stages linked to children's ages that were connected to learning skills. In the table below, the four developmental stages are listed, and the core reasoning skills are listed below.

Table 1: Jean Piaget's Four Stages of Development (Piaget & Cook, 1977; Santrock, 2010, 2013)

Age	Development Stage	Acquired Reasoning Skills
Birth to 2 years	Sensorimotor	Reflexive to symbolic thought
2 to 7 years	Preoperational	Symbolic thought improves with word usage
7 to 11 years	Concrete Operational	Logical reasoning of concrete aspects
11 years to adulthood	Formal Operational	Reasoning becomes abstract and more logical

The cognitive growth of individuals in these stages involves an interchange between biological and environmental factors. The processes whereby this occurs are adaptation, disequilibrium, and the developmental stages (Piaget & Cook, 1977; Santrock, 2010, 2013).

According to the theory, from age 0 to 2 years, the baby learns to distinguish between objects. He or she uses memory and thinking and becomes more active. From age 2 to 7 years, the child can communicate using language and can employ transductive reasoning. The child is also egocentric in their thinking. From age 7 to 11 years, the child can distinguish between ideas and engage in deductive reasoning. The child is, however, unable to reason abstractly at this point (Piaget & Cook, 1977; Santrock, 2010, 2013).

From 11 years to adulthood, the formal operational stage, the individual demonstrates the ability to reason abstractly and engages in hypothesis testing. The thinking at this stage becomes scientific, and problem-solving skills become more advanced (Piaget & Cook, 1977; Santrock, 2010, 2013).

The stage of interest for this study is the formal operational stage, which is focused on adult reasoning. This stage is important as it represents the stage in which individuals utilise logical abstract reasoning. This stage also involves hypothetical-deductive reasoning, which is a problem-solving technique that individuals develop. This technique allows them to make presumptions about a particular problem and logically address the problem (Piaget & Cook, 1977; Santrock, 2010).

When individuals use concrete operational thought, they tend to make duplicate mistakes, because they cannot yet learn from their errors. Additionally, when adults have difficulty with formal operational thought, they are in a transitional phase, which implies that they cannot consistently use formal operational thought. As a result, formal operational thought allows individuals to solve problems and prevents them from repeating mistakes (Wankat & Oreovicz, 1993).

Intelligence measures aim to assess the schemas (which refer to how individuals think) that individuals use. This also includes processes such as assimilation and accommodation. The transition across (the interplay between) schemas are known as the equilibration process, which includes equilibrium and disequilibrium. Equilibrium occurs when there is no conflict between the schema and the experiences of the individual, while disequilibrium refers to schemas that are in conflict with an experience; and essentially the equilibration process can also be referred to as a learning process (Piaget & Cook, 1977; Santrock, 2010, 2013).

A process referred to as “groping” occurs when an individual attempts to either form a new schema or change an existing schema. This groping allows a new equilibrium to occur. It should also be noted that groping allows optimal learning to occur, while challenges (either too great or too low) will cause the individual to either become bored or withdraw (Piaget & Cook, 1977).

Furthermore, Piaget viewed equilibrium as the harmony between the processes of assimilation and accommodation. Assimilation refers to individuals acquiring new information and merging this into an existing schema of knowledge. Accommodation, on the other hand, refers to the changing of an existing schema due to new information having being acquired. These concepts are important when trying to understand how individuals incorporate new information into their schemas of knowledge. Additionally, particular types of thinking are necessary for individuals to advance cognitively through the developmental stages. If individuals have not progressed to abstract thinking, they will encounter problems academically when situations requiring such thought are necessary (Blake & Pope, 2008; Piaget & Cook, 1977; Santrock, 2010). It should be noted that the concept of equilibrium relates to Rasch’s statement on attainment.

The processing of information, therefore, occurs when the information provided is not contradictory to what the individual already knows, which allows it to be integrated into existing knowledge. This process is also referred to as accommodation. When new information is obtained that differs from the existing structure of knowledge, it is excluded, assimilated, or transformed. Individuals who are using concrete operational thought would struggle to assimilate and would then either discard the information or use a strategy of memorising without understanding. People who can use formal operational thought would use strategies of assimilation and transform the information with understanding (Piaget & Cook, 1977; Wankat & Oreovicz, 1993).

Criticisms levelled at Piaget's theory were based primarily on the age-defined stages, arguing that these ages were not always accurate in suggesting the type of thought that children were engaging in. The influence of culture and education are instrumental in developing children's thinking, and this was not acknowledged by Piaget's theory. His theory has however been imperative in allowing one to understand the cognitive development of children (Santrock, 2010).

It should, however, be noted that the thinking strategies he theorised about as well as the use of abstract reasoning are important considerations in the development and interpretation of the individual's performance on the ECT.

3.2.2 Lev Vygotsky Socio-Cultural Theory of Cognitive Development

Lev Vygotsky had a vastly different approach from Piaget to understanding the intellectual functioning of individuals since he was influenced by Marxist theory. Vygotsky believed that social institutions and interactions assisted in the cognitive development of individuals (Vygotsky, 1978). He believed that the importance of the socio-cultural context could not be underestimated and was crucial for furthering intellectual development. Vygotsky identified cognitive learning zones such as the zone of actual development and the zone of proximal development. The zone of actual development referred to individuals having the ability to complete tasks by themselves, without assistance. This zone consisted of tasks that required only their present ability and hence they were not learning anything new in their completion of the tasks. As a result, Vygotsky endorsed the zone of proximal development, because he believed that tasks in this zone challenged individuals, and would provide for optimal cognitive development. These were tasks that individuals needed assistance with as they were unable to complete them on their own (Blake & Pope, 2008; Ormrod, 2008; Santrock, 2010; Taylor, 1994; Vygotsky, 1978).

Vygotsky emphasised the impact of culture and the role of language on individual cognitive development. He believed that the development of language was an inclination that individuals had from birth. Language develops in a systematic process, in which infants learn words and articulate their words at around age 1. Toddlers at approximately age 2 form sentences and the complexity of these sentences expand until they are in pre-school. When children reach age 5 and 6, their ability to use language becomes similar to adults (Ormrod, 2008; Vygotsky, 1978).

Vygotsky identified the concept of scaffolding, which refers to a child being assisted by an adult when attempting to use problem-solving skills. Scaffolding makes use of learning and experiences. This process of scaffolding also allows individuals to improve their reading skills and consequently, Vygotsky's theory was influential in programmes such as Reading-and-Recovery and Guided Reading (Blake & Pope, 2008; Santrock, 2010, 2013; Vygotsky, 1978).

The most important considerations of Vygotsky's theory are his recognition of social and cultural factors that influence learning and development. This is significant in light of the multicultural context in which the ECT was used.

3.2.3 A Comparison of Piaget and Vygotsky's Theories of Cognitive Development

Piaget identified stages that individuals had to complete or achieve, while Vygotsky believed that individuals needed environments not above or below their abilities to stimulate learning. Piaget believed that experiences allowed individuals to learn, yet he also argued that cognition was not affected by language skills; hence, cognition preceded the development of language. Vygotsky, however, believed that language was essential in learning, especially when individuals needed to complete demanding tasks (Blake & Pope, 2008).

Piaget viewed learning to take place on an individual basis, dependent on the individual. Vygotsky, however, believed that socially interacting furthered the development of intellectual and language abilities. As a result, individuals developed their skills and abilities in their social environments. Piaget's theory encouraged individuals to be the source of their own learning, while Vygotsky's theory emphasised that individuals learn from their social environment, thus the environment is the source of learning. Piaget believed learning was internalised while Vygotsky externalised the learning experience (Blake & Pope, 2008).

Even though Piaget and Vygotsky differed in their construction of theories, they both believed that learning was needed for higher-order thinking to occur. These theories assisted educators in understanding individual's learning processes and why some struggle with academics while others excel (Blake & Pope, 2008). These theories highlight the issues related to testing intelligence in individuals as well as the implications considered in cross-cultural testing.

Piaget and Vygotsky were instrumental in formulating an understanding of cognitive functioning for children and adults. Moreover, their contrasting views allow for a comprehensive depiction of cognitive functioning in humans. Their theories, however, fell short of adequately exploring the cognitive functioning of adults, as Piaget's stages were predominately focused on children and did not allow the adult perspective to be fully understood. For this reason, it is essential to explore the theories that endeavoured to concentrate on adult cognitive functioning. The three theories that will be explored as part of broadly discussing the cognition of adults are Schaie and Willis' staged theory of cognition for adulthood (2000); Perry's theory of the development of college students (1970); and

Belenky, Clinchy, Goldberger, Nancy and Tarule’s theory of Woman’s ways of Knowing (1986).

3.2.4 Schaie and Willis’s Staged Theory of Cognition for Adulthood

Theorists such as Erik Erikson, who theorised about psychosocial development, and Paul Baltes’s theory on selection, optimisation, and compensation only attempted to broadly address the development of individuals across the lifespan. There was, however, a need for a more comprehensive theory that pertained to the psychological advancement of individuals, with a specific focus on adulthood. According to Erikson’s psychosocial model, individuals progress through psychosocial stages that require them to resolve various conflicts from birth to their eventual old age and death. Baltes’s selection, optimisation, and compensation theory also poses a staged approach to individuals aging with each stage having different psychological achievements and failures. The influence of society tends to decrease as individuals age (Giri, 2003; Schaie & Willis, 1993, 2000; Schaie, 2008).

To address this shortage of developmental theories focusing specifically on adulthood, Schaie and Willis proposed a stage theory of cognition (2000) addressing developmental theories in adulthood. This theory was developed based on research focused on the cognitive development of adults. The theory proposes seven stages through which individual’s progress (Schaie & Willis, 1993, 2000; Schaie, 2008).

The first stage is labelled acquisitive since it relies on much of Piaget’s theory of children. Consequently, this stage refers to the period of childhood and adolescence.

The second stage covering young adulthood is labelled achieving. This stage involves individuals applying the knowledge they acquired in the previous stage. This knowledge is used for embarking on career opportunities and other life changes such as marriage and children. This stage also involves the use of intelligence and goal-orientated behaviour.

The third stage is labelled responsible, which is indicative of the social context of the individual. This also refers to tasks and decisions made in family contexts. As a result, individuals learn to assume roles of responsibility in family and work situations.

The fourth stage can be achieved if individuals have appropriate opportunities to reach a higher level of responsibility than the third stage. The level of skills and knowledge the individual is exposed to through their work environment enables them to arrive at this fourth stage, which is labelled the executive stage.

The fifth stage is called the reorganizational stage. It involves intellectual functioning at a high level, as individuals are beginning to consider decisions that will impact their retirement. This stage occurs in the period ‘young-old’.

The sixth stage is labelled re-integrative, because the knowledge people use is based on hobbies. This stage requires individuals to adjust their use of knowledge in a different way. They are required to make faster decisions as time is no longer a luxury. This also includes decisions about their assets and testament, as death is a reality.

The last stage is labelled legacy creating. Depending on the career of the individual, they will retain certain levels of cognitive skills. This period is usually used for individuals to put their affairs in order and distribute valued items among family members.

3.2.5 Perry’s Theory of the Development of College Students

Another theory of cognitive development pertaining to adults is Perry’s theory of the development of college students (1970). He studied the behaviour and thinking of Harvard University students for the duration of their degree courses (four years) and this allowed him to formulate a theory on college students. The goals of his theory rested on two ideas, namely a college student’s transition from a dualistic to a relativistic view of the world, and college students' ability to commit to this relativistic worldview (Perry, 1970; Wankat & Oreovicz, 1993).

Fundamental to Perry's theory is the ability of college students to evolve, which requires learning to have taken place. Thus, if they are not in a particular space for learning to take place, then they will not be able to complete the tasks required. Perry's theory proposed nine positions through which college students progressed, namely (Perry, 1970; Wankat & Oreovicz, 1993):

Basic Duality: This stage involves the college students viewing things as either correct or incorrect; there are no grey areas (no other options).

Dualism - Multiplicity Pre-legitimate: This stage involves the awareness of multiplicity (several options), yet the college student is still engaging in dualistic thinking.

Multiplicity Subordinate or Early Multiplicity: This stage involves the college student accepting that multiplicity is inescapable, and that knowledge can be unclear at times.

Complex Dualism and Advanced Multiplicity: At this stage, the college student realises that dualistic thinking cannot continue and this allows them to either accept this and become autonomous thinkers or break away from this. College students choosing to break away entails that they reject the idea that dualistic thinking cannot continue and so continue to think dualistically.

Relativism: This stage involves the college student changing their perception of the world as they now begin to view it as relative, since there are not always absolutes.

Relativism - Commitment Foreseen: This stage involves the college student accepting their changed view of the world, and they become more independent in their style and identity.

Commitment of own Free Will: This stage involves the college student pledging to do things of their own free will.

Stylistic Issues of Commitment: This stage involves the college student making commitments towards future careers, marriage, children and so forth. .

Maturity Associated with Commitment and Styles: This stage involves the college students' confidence in committing to their own style and thinking.

These positions were guided by three phases, namely temporising, retreating, or escaping. Temporising refers to a break in the learning from one phase to another, while retreating refers to instances when individuals move back to a previous position. Escaping, on the other hand, involves an individual resisting commitment (Wankat & Oreovicz, 1993).

3.2.6 Belenky, Clinchy, Goldberg, and Tarule's Theory on Woman's Ways of Knowing

Perry's college-staged theory (1970) was, however, criticised because it only focused on males at university and the methodology had several limitations. For this reason, a female perspective was required as their cognitive development would be different from males. Belenky, Clinchy, Goldberg, and Tarule (1986) replicated a similar study on females, but this sample was not only confined to female college students as in Perry's case of only male college students, as it included other females as well. Belenky et al. (1986) identified seven positions through which women progress in their journey of acquiring new information (Garrison, 2009).

Position 1 is labelled Silence. This stage involves women being quieter than males as they have had generally speaking never been in positions to voice their opinion. Women were never fully acknowledged or asked for their opinions, so they did not have the freedom to share their thoughts. Thus, women require time during this period, as they must now learn to interact socially and share their thoughts with others.

Position 2 is labelled Received Knowledge. This period is characterised by the accumulation of knowledge. The woman seeks to retain all information from others. She starts to engage socially and has a need to know more.

Position 3 is labelled Subjective Knowledge. This stage is where the woman becomes more conscious of her opinions and thoughts towards things. She starts to realise her worth and that she may have opinions. This realisation opens her up to forming her opinion on matters. Her way of thinking shifts and she begins to want to interact in discussions.

Position 4 is labelled Quest for Self. This is when the woman acknowledges her voice and thoughts. She feels more comfortable engaging in male spaces. This new interaction with information allows her to engage with ideas that conflict with previous sets of knowledge. She then explores these new ideas, but holds onto her responsibilities.

Position 5 is labelled Procedural Knowledge. This position is viewed as the woman acquiring reason through a process of learning. This represents the woman being able to shift through all the different views and ideas. She is more aware of herself and her beliefs. Her cognitive skills are more advanced and decision-making ability is improved.

Position 6 is labelled Separate and Connected Knowing. The woman is able to distinguish between different forms of knowledge. She stays away from negative influences. Her knowledge is settled deep within her and she will resist being silent.

Position 7 is labelled Constructed Knowledge. The woman is actively involved with debates and knowledge production. Her development is both cognitive and moral as she becomes confident in her knowledge.

This theory is significant as it attempts to categorise the different thinking processes in which individuals in a college environment engage. Since the age group of the ECT falls within typical university attending ages, it is worth considering.

In addition to the theories of Schaie and Willis (2000), Perry (1970) and Belenky et al. (1986), there are other theories that assist in shaping how adult cognition is understood. One of these theories is the co-constructive perspective that contributes to comprehensively understanding adult cognitive functioning. The co-constructive perspective is a theory that combines biology and culture. This view is comprised of three fundamentals. The first is that there are advantages associated with the biological changes that occur in early years compared to those in later years. This refers to the progressive selection processes that increase development in individuals. The second refers to instances when cultural resources multiply, allowing the development of individuals to be increased. This is essential to individuals in their later years as it assists in the aging process. The third refers to the neurobiological functions that decrease in older individuals, which affects the advancement of cultural resources for individuals in later years (Li, 2003; Perry, 1970; Schaie & Willis, 2000; Schaie, 2008).

The process of co-construction involves interactions in a social setting with individuals of influence, such as teachers, peers, or parents. Secondly, it is an engagement with others to arrive at a mutual conclusion. This engagement requires combined understanding and negotiation to argue different points of view. Thirdly, it is the result of the engagement that involves mutually viewed individual and social cognitions. This involves academic task completion, motivations, and conceptual thinking. There are various perspectives within the co-construction viewpoint, such as neo-Piagetian, neo-Vygotskyan, and situated and socially shared cognition. These differing perspectives indicate how cognitive development occurs in a variety of ways (Li, 2003; Reusser, 2001; Willis & Schaie, 2006).

From the Piagetian perspective, cognition is partially as a result of individual processes and the acceptance of other individuals' perspectives to incorporate other perspectives. The Vygotskyan context argues that cognitive development is reliant on processes involving culture and society and does not include the individual's participation. Thus, learning is external to the individual. The situated and socially shared cognition view sees learning as intrinsically part of socio-cultural environments. This is seen when individuals engage in a discussion of which the result is new to both, as the solution was formed through the collaboration (Reusser, 2001).

These theories and views are significant in understanding the complex nature of human development. For one to consider the intellectual functioning of adults in assessments, one must be aware of the different ways in which they may acquire intelligence. Cognitive development over a lifespan is guided and influenced by many biological factors, social activities, and environments. The identification of these aspects will consequently create a comprehensive view of cognitive thinking skills in adults.

3.3 Models of Intelligence

Historically, there are two approaches to understanding intelligence, namely a universal "g" factor and multiple factors approach. Essentially, the universal "g" factor approach is demonstrated by theorists such as Cattell (1963) who posited a general factor of intelligence as well as sub-factors making up this composite construct. The multiple factors approach which include theorists such as Gardner (1983) who denied the existence of a general intelligence factor and argued for the existence of multiple factors that equally exist to comprise intelligence (Pal, Pal, & Tourani, 2004).

The earliest theory defining intelligence was the faculty theory (18^{th} – 19^{th} century). As its name implies, it was theorised that intelligence consisted of different faculties, such a reasoning, memory, and imagination, which functioned independently of each other. These various faculties required training to improve, but remained unrelated to each other. This theory did not withstand criticism and was hence replaced by other theorists. There was a short-lived theory, known as the one-factor theory, which suggested that intelligence was comprised of different elements to form a common factor. This meant that all individuals had varying levels of abilities within their intelligence (Pal, Pal, & Tourani, 2004).

3.3.1 Theories of a General (Unifying) Intelligence

The structural approach to intelligence theories used techniques such as correlations and factor analysis to identify the construct of intelligence. This approach was responsible for defining how intelligence was understood and how tests were constructed. Theories guided by this approach were the following: sensory response theory (Galton, 1883), intelligence quotient theory (Binet & Simon, 1908), Spearman's two-factor theory (1904), Thurstone's primary mental abilities (1938), and hierarchical theories (Pelser, 2009).

Galton (1883) initiated the idea that intelligence could be a hereditary consequence and that individuals in certain family lineages were born intelligent. He coined the term "hereditary genius", as intelligent individuals were found in families that had a history of intelligent people. This theory was however not able to withstand scrutiny, and this concept was followed by other theorists who delved into understanding intelligence. Galton was, however, recognised for his contribution to statistical procedures for analysing relationships between variables, known as correlations (Galton, 1883; Sternberg, Jarvin, & Grigorenko, 2011; Taylor, 1994). Galton suggested that intelligence is an ability acquired by sensory functions since it is the means by which individuals acquire information. Galton and Cattell extended the theory and constructed mental tests to assess individual's sensory functions. This involved perceptions and psychomotor abilities and conceptualised the notion of a general mental ability. The sensory responses did not provide evidence to sustain the existence of the theory (Pelser, 2009).

The intelligence quotient theory by Binet and Simon (1908) was proposed to define intelligence. Their theory was based on the conception of higher-order functions such as reasoning and knowledge, while lower-order processes such as sensory functions were not of concern. This influenced how intelligence measures were created (Binet & Simon, 1908; Pelser, 2009; Sternberg et al., 2011).

Binet, in France, created a measure by which to assess children's intelligence. He did so by indicating that it would measure their existing ability and theorised that their ability would increase as they aged. He also hypothesised that intelligence was an accumulation of different abilities that would increase with age. For this reason, he created mental age norms. This test was then adapted in the USA and became labelled as the Stanford-Binet test and was used internationally. This test was, however, used as exclusion (in a discriminatory way) measure for immigrants and within the USA military (Brown, 2016; Pelser, 2009; Sternberg et al., 2011).

Spearman, in his theory of intelligence (1904), was one of the early theorists who assisted in formulating a theory of the understanding of intelligence. He used factor analysis to dissect the different components he believed were associated with intelligence. This theory was the first of its kind, and the correlations between the different variables indicated that they shared an underlying factor. This was the conception of "g", as a general factor. Spearman hypothesised that "g" could be both mental speed and working memory or mental self-government, as this would explain what was common to all the tests he analysed (Almeida et al., 2011; Spearman, 1904; Sternberg et al., 2011; Taylor, 1994). The conceptualisation of “g” led to Spearman’s two-factor theory. The two factors he identified were “g” and “s”, which were labelled general ability and specific ability respectively. The factor “g” was thought to be innate ability and the more of the “g” ability that one had, the more successful one would be. The factor “s” referred to abilities acquired in one’s environment and would be different for different people depending on their surroundings (Bekwa, 2016; Pal, Pal & Tourani, 2004; Spearman, 1904; Sternberg et al., 2011).

Spearman’s theory was in accordance with the concept of the general mental ability. The factor “g” was common to all cognitive assessments, while factor “s” was exclusive to tests. Since factor analyses of cognitive tests were fundamental to his theory, speculation of “g” as an actual factor or a product of analysis was debated (Pelser, 2009).

Cattell (1963) also analysed the information and then conceptualised “g” as general intelligence with two core factors, fluid and crystallized intelligence, hence it was named the two-factor theory (Bekwa, 2016; Brown, 2016; Cattell, 1963; Horn & Cattell, 1966; Marshalek, Lohman & Snow, 1983; Schaie, 2008; Sternberg et al., 2011). This dual intelligence model involved the development of crystallized intelligence being impacted by fluid intelligence. The later part of adulthood is the period in which crystallized intelligence is utilised as the accrued aspects such as education are recognised, and engaging in cognitive stimulation assists in preserving it. Fluid intelligence is affected by chronic diseases such as high blood pressure and biomarkers due to its neurobiological influences. As a result, crystallized intelligence methods are used to assist with the damages associated with fluid intelligence (Bekwa, 2016; Brown, 2016; Griskevica & Rascevska, 2009; Schaie, 1993, 2006, 2008; Taylor, 1994).

In line with this is the investment hypothesis of intelligence by Cattell, which promotes the idea that cognitive abilities are subject to environmental, genetic, and learning opportunities (Van der Pool & Catano, 2008). The impact of age on constructs such as fluid intelligence and crystallized intelligence is rather interesting. The construct of fluid intelligence is at its best for individuals aged 22, while crystallized intelligence peaks for individuals at the age of 36. An individual's general cognitive abilities are at their best at age 26, while they start to decline at age 52. Additionally, fluid intelligence stabilises between ages 18 and 28, while crystallized intelligence can increase over all ages. It is worth noting that although cognitive processing speed is related to the decline in age for memory and spatial ability, it, however, does not hold true for verbal ability (Griskevica & Rascevska, 2009; Schaie, 2008).

In a longitudinal study over a period of seven years that studied individuals aged between 25 and 88 years, it was found that there was an asymptotic result for word fluency, which was achieved by age 39, while inductive reasoning and verbal ability was achieved by age 53. When observing how individuals performed over time, there was a direct increase for inductive reasoning and spatial orientation, while verbal ability obtained a peak and then declined gradually. Interestingly, word fluency declined to a certain point and then gradually increased to a particular point. This reflects the impact of age, which has been referred to ageism because old age is associated with a decrease in cognitive ability and unwanted psychological problems become evident (Schaie, 2006, 2008). The results of this study are intriguing in light of Cattell's theory of fluid and crystallized intelligence.

Vernon (1950) developed a hierarchical theory which was a combination of Spearman and Thurstone's theory. He identified levels of intelligence and abilities that would differ depending on the level. The highest level consisted of general intelligence. Vernon's hierarchical model proposed "g" at the head and was sub-divided into verbal-education and practical-mechanical. The subsequent level comprised of verbal-numerical and educational factors, practical-mechanical, and spatial-physical abilities. The successive levels included factors that consisted of the above factors. This sub-divided into verbal and numerical ability, spatial, and mechanical ability. Within these factors, there are more specific factors. These specific factors were the lowest level. Vernon was guided by the notion that environments and genetics could have an impact on intelligence and mental abilities. He found that 60% of the differences that were observed between individuals on intelligence measures were due to genetics. He also noted that genetic differences assisted in explaining racial differences regarding people of different race's mental abilities (Marshalek et al., 1983; Pal, Pal, & Tourani, 2004; Pelser, 2009; Sternberg et al., 2011; Taylor, 1994; Vernon, 1950).

Anderson's theory of cognitive development (1992) suggested that regardless of how individuals are constructed, the process whereby they were able to adapt to situations and solve problems allowed them to function at their prime. He referred to this process as rational analysis. This form of analysis involved considering all elements in the surrounding, identifying objectives, considering arguments that were linked to cognitive processes, and creating the best response to the problem (Anderson, 1992; Pal et al., 2004).

Eysenck's structural theory (1973) consisted of three neurological processes, namely reaction time, inspection time, and average evoked potential. The first two processes referred to observable phenomena, while the third process was comprised of mental waves in the brain. He argued that the more intelligent the individual, the less time they would take to respond, and the more intricate their mental waves would be (this was observed by an electroencephalogram) (Eysenck, 1973; Pal et al., 2004).

The combination of a hierarchical system of intelligence and the inclusion of the "g" factor extended the theory of intelligence by Cattell. The different cognitive abilities associated with intelligence formed a pyramid-like structure. This hierarchical model was referred to as the Carrol-Horn-Cattell (CHC) model, which identified three strata of intelligence. Situated at the top was "g" (which is identified as general mental ability). Below "g" are factors such as gf (fluid ability), gc (verbal crystallized ability), gv (spatial visual ability), and gm (memory ability). The last level contains factors such as verbal comprehension, verbal fluency, inductive reasoning, spatial visualisation, and perceptual speed. This last level contains concepts that are more psychologically visible in assessments than the higher up factors, because the broad nature of the higher up factors allows them to be associated with many elements (Almeida et al., 2011; Lohman & Lakin, 2009; Sternberg et al., 2011).

This model identified two factors within the second stratum that are of particular concern for this study. The first factor, known as fluid intelligence (Gf) is the ability to use inductive and deductive reasoning to solve complex problems, while the second factor is crystallized intelligence (Gc), which consists of verbal knowledge and skills, and involves language, culture, education, and experience (Horn & Cattell, 1966; Kvist & Gustafsson, 2007; Marshalek et al., 1983).

These cognitive abilities are interdependent, and the abilities relating to identifying numbers and words stem from inductive reasoning skills. Research concerning cognitive abilities is aligned with the Gf-Gc theory. It was theorised that education and culture impact crystallized ability, while fluid ability is associated with individual and learning occasions (Brown, 2016; Marshalek, 1981). This theory influenced the development of many contemporary psychological tests tapping into intelligence. The prevailing view on cognitive assessment thus makes use of the combination of fluid and crystallized intelligence.

3.3.2 Multiple Intelligence Theories

The theories that opposed the view of a unified intelligence (g) were Gardner (1983), Sternberg (1988), Thorndike (1927), Thurstone (1938), Guilford (1956), and Ceci’s (1990) theory of multiple factors (Almeida et al., 2011; Pal et al., 2004; Sternberg et al., 2011).

Gardner’s theory of multiple intelligences (1983) was based on the idea that several forms of intelligence existed in individuals. This meant that a person’s strengths were his or her areas of intelligence. For example, linguistically strong individuals would be gifted in linguistic intelligence, and careers that require linguistically strong individuals would be best suited to these individuals. The eight forms of intelligences are the following: linguistic, logico-mathematical, musical, spatial, bodily-kinaesthetic, naturalist, interpersonal, and intrapersonal (Almeida et al., 2011; Gardner, 1983; Pal et al., 2004; Sternberg et al., 2011).

Since Gardner believed in multiple intelligences, he denied the existence of a general intelligence factor. To cement his theory, he utilised neurological proof to support his notion that abilities were independent of each other. He demonstrated that when certain areas of the brain were harmed, other parts of the brain functioned and became strengths. Linguistic intelligence, which is of particular importance for this study, was found to be one of the two forms of intelligence found in intelligence tests. This was evident in the factor analysis and substantiated the existence of this form of intelligence (Gardner, 1983; Pal et al., 2004).

This led to Gardner suggesting that all eight factors have equivalent status. His theory received criticism, and soon, other theorists endeavoured to supply answers. Gardener argued against how intelligence tests were constructed and that other types of intelligence were not considered in the construction of tests. As a result, the theory of multiple intelligences led to the creation of tests like the General Aptitude Test Battery (Pal et al., 2004).

Sternberg’s triachic theory (1988) was also built on the foundation of multiple intelligences, but he condensed them into three. The first, componential intelligence, served to explain analytical abilities, problem-solving abilities, and academic inclinations. The second, experiential intelligence, was comprised of artistic and creative abilities. This included creative intelligence, which involves adapting to various situations and experiences influencing the capacity to solve new problems more quickly. The third, contextual intelligences, consisted of solving common daily problems and were associated with practical understanding and ability. This practical intelligence involved adapting to daily situations (Bekwa, 2016; Pal et al., 2004; Sternberg, 1988; Sternberg et al., 2011).

Sternberg believed that there were two ways of defining intelligence: by operational or real definitions. The operational definition of intelligence was one that could be measured and was visible (statistically). The real definition of intelligence was the essence of what comprised intelligence (Pal et al., 2004).

Thorndike’s multi-factor theory (1927) was in opposition to a general ability. He identified with Spearman and conceived his theory on the basis that individuals have various attributes associated with their abilities. He identified four aspects of intelligence, namely the level of difficulty, the range of tasks, the area or magnitude of aspects that individuals can respond to, and the speed at which individuals respond. Thorndike stressed that intelligent abilities were the manifestation of vastly different factors (Thorndike, 1927; Pal et al., 2004).

Thurstone introduced the theory of primary mental abilities (1938), or group factor theory. Thurstone’s theory was in opposition to Thorndike’s theory and the idea of a general factor of intelligence. His theory was based on his analysis of intelligence tests, which caused him to argue against a general intelligence factor. Among the primary mental abilities, he identified verbal relations, word, memory, inductive reasoning, and deductive reasoning. One could argue that these five categories form part of verbal reasoning. Thurstone also ran a factor analysis to understand intelligence, and found seven factors (Bekwa, 2016; Marshalek et al., 1983; Sternberg et al., 2011; Taylor, 1994; Thurstone, 1938).

He identified primary factors such as number, verbal, space, memory, word fluency, and reasoning. Based on these factors, he constructed the Test of Primary Mental Abilities. He believed (aligned to Spearman’s theory) that there are many mental abilities that are composed of primary factors with their own functions and use. These primary factors are independent, as they are psychologically and operationally different. These mental abilities were clustered into 11 categories, which were theorised to form an individual's intelligence (Marshalek et al., 1983; Sternberg et al., 2011; Taylor, 1994; Thurstone, 1938).

Guilford's model (1956) of the structure of intellect identified a three-structure theory of intelligence, namely content, mental operations, and operational products. Within the content structure, the following aspects were found: visual, auditory, symbolic, semantic, and behavioural. The mental operations included: cognition, memory retention, memory recording, divergent production, convergent production, and evaluation. The operational products consisted of units, classes, relations, systems, transformation, and implications (Guilford, 1956; Marshalek et al., 1983; Pal et al., 2004; Sternberg et al., 2011).

Ceci's bio-ecological theory (1990) suggested that individuals had multiple forms of cognitive potential. These cognitive potentials were biologically determined, were impacted by cognitive abilities, and were influenced by the environment of the individuals. Factors such as personality, motivation, and education were known to affect cognitive abilities. The circumstances as well as the mental, social and physical conditions of the individual have an influence on these abilities (Ceci, 1990; Pal et al., 2004; Sternberg et al., 2011).

These opposing theories are instrumental in considering how intelligence measures were and are created. It is, however, interesting to note that verbal ability was identified in both streams of thought as a component of intelligence. This allows one to be certain of its existence regardless of the system of thought.

3.3.3 The Influence of the Intelligence Models in Testing

Jensen (1974) argued that measuring "g" be inevitable when assessing ability in tests because the factor "g" requires individuals to reason, and reasoning forms part of "g".

Therefore, “g” is a definite component of intelligence tests. Jensen also noted that “g” was widespread to numerous ability tests and is related to learning potential measures. The “g” factor is strongly related to job performance, and the validity of such measures (containing “g”) is usually reasonable, and never approaches reliability coefficients of 0. Specialised careers are selected with the use of intelligence measures. Examples of these are aviation selection and pilot training (Jensen, 1974; Hunter, 1986; Pelser, 2009). Fagan (2000) argued that defining intelligence as a form of processing would limit the effect of culture on intelligence assessments as well as limiting other unrelated factors influencing performance on intelligence measures.

Gignac (2006) argued that crystallized intelligence was a good indicator of general intelligence (g), and he justified his argument regarding the verbal subtests (vocabulary and information) of the Wechsler tests having higher loadings on the main factor (intelligence). This allowed him to conclude that verbal tests are the best indicators of general intelligence (Kvist & Gustafsson, 2007). Other related research conducted on the crystallized intelligence factor in assessments has associated this factor with verbal tests, because of its relation to academic achievement. This model has been used to evaluate cognitive ability and has been connected to comprehension, yet the way in which it measures comprehension refers to the quantity and not quality of the text. Essentially, individuals obtaining high crystallized intelligence scores may exhibit a broad knowledge of many areas but may lack in the depth of their understanding (Horn & McArdle, 2007). This translation of how crystallized intelligence scores are interpreted differs to the historical intents originated by Cattell (Horn & McArdle, 2007) for crystallized intelligence.

The CHC model has influenced the development of several measures of intelligence and has been instrumental in the advancement of cognitive testing. Numerous cognitive tests predominantly concentrate on measuring one aspect of intelligence, which does not accommodate the requirements of psychologists for selection, development, or diagnostic purposes. Multiple assessment measures therefore need to be used so that different cognitive functions can be evaluated to form a comprehensive cognitive summary of the individual (Flanagan & McGrew, 1997). Moreover, cognitive ability tests are known for their utility in selection and decision-making and are usually good predictors of ability (Koczwara et al., 2012). Moreover, theories on intelligence models are necessary for the development of cognitive measures, as they have assisted in explaining how the individual functions cognitively.

Although the Gf factor involves the process of inductive reasoning, it has also been compared to deductive reasoning. Inductive reasoning has also been described as a cognitive process by which knowledge is gained, and can be used in different and new situations. This suggests that there is a strong relation between inductive reasoning and the use of knowledge (Csapo, 1997; Keeves, 1992). More so, studies have found links between intelligence (g), fluid ability (gf), and inductive reasoning. This relationship is known to be a good predictor of academic achievement and allows one to conclude that effective cognitive skills are reliant on good reasoning skills (Lohman & Lakin, 2009).

Studies exploring the psychological concept of inductive reasoning have linked it to critical thinking, creative thinking, hypothesis testing, and the development of concepts. Development in educational settings can also aid inductive reasoning; specifically school instruction can aid the development of inductive reasoning in young individuals, since one’s younger years are the most critical time in which inductive reasoning develops (Keeves, 1992). If one were to consider the implication of these findings, it would imply that older individuals have less chance of developing their ability to inductively reason. This means that adults being tested have already developed their inductive reasoning skills.

A large part of fluid intelligence consists of working memory, and thus part of understanding cognitive ability involves an understanding of working memory. Cognitive processes such as encoding and retrieving of images form part of the functions of working memory. Reading is comprised of distracter tasks (processing tasks), which interrupt the encoding of information. Reading is guided by the time-based resource-sharing theory, which explains the relation of decaying to attentional refreshing. Decay refers to forgetting and is explained as interruptions in the encoding and retrieval of information processes. Attention refreshing relates to repairing working memory by guiding attention to certain tasks. The theory is centred on the idea that with time, memory starts to decay (Oberaer & Lewandowsky, 2013).

The decaying of memory is avoided by using the mechanism of refreshing. The process of refreshing occurs in the instances when the distracter tasks are not used. The performance of memory is the time shared between the distracter tasks and refreshing, as only one of these processes can occupy attention at a time. The refreshing process occurs during pauses, such as the time between reading two words. Within this process, it is vital to reflect on the amount of time that the distracter task is receiving attention with no refreshing; this is usually referred to as cognitive load. More importantly, all of these processes occur within seconds (Oberaer & Lewandowsky, 2013).

A hindrance to the validity of cognitive testing, especially when intelligence is assessed, is the effect of speed on intelligence assessments. It may assist in differentiating high performing from low performing individuals, but it can, however, contaminate the data and interfere with the construct validity of the assessment (Keith & Reynolds, 2010). Additionally, Oberaer and Lewandowsky (2013) found that the time pressure associated with distracter tasks affected memory retention. This means that assessments with time pressures may be measuring aspects of working memory instead of the intended construct. This interpretation was based on the cognitive control and verbal representations used to provide faster responses before the time lapsed (Oberaer & Lewandowsky, 2013). Time is a significant variable to consider, as it often threatens the validity of cognitive instruments as the construct being measured is compromised. This aspect was taken into account with the second piloting of the ECT as the time limit could have affected how individuals performed on the test. The current version (ECT version 1.3) does not have a time limit, and so the issue of speed was removed.

Some tests, such as vocabulary tests, could be overlooked for cognitive performance, but studies have shown that there is a relationship between vocabulary and reasoning, because vocabulary uses cognitive processes such as drawing inferences and comprehending (Marshalek, 1981). This finding suggests that vocabulary tests could contain essential reasoning skills that allow for better academic performance by individuals. Hence, reasoning is considered as one of the most important psychological constructs because it is instrumental in learning, education, language advancement, and performance on assessments (Lohman & Lakin, 2009).

It was found that verbal short-term memory is vital in the acquisition of vocabulary for local and foreign language speakers. Additionally, it was established that bilinguals knew fewer words than monolinguals and had a reduced vocabulary efficiency, which was explained by bilingual’s exposure and acquisition of words in multiple languages (Engel de Abreu, Baldassi, Puglisi, & Befi-lopes, 2013). Furthermore, vocabulary assessment is associated with crystallized intelligence (Cockcroft, Bloch, & Moolla, 2016).

In industrial and educational psychology practices, the use of tests for selection purposes are predictable and an essential method of assessing aptitude. Selection decisions were, however, hampered by the apprehension caused by research conducted on cognitive ability tests across races because it indicated that Black individuals consistently scored lower than White individuals. This was observed in tests such as the Senior Aptitude Test (SAT) and the Wechsler Adult Intelligence Scale (WAIS). In a study on Amerindian's (Native Americans) performance on cognitive tests, it was suggested that the reason they scored lower on verbal tests was because it was not in their native tongue. It was observed; however, that they scored better on non-verbal tests, as reading and writing was limited in these tests (Van der Pool & Catano, 2008). Moreover, cognitive ability tests are known for their utility in selection decision making and are generally good predictors of ability (Koczwara et al., 2012).

3.4 Language and Cognition

"It seems to me that, in so far as we understand anything about cognition…we discover very specific mental structures developing in the course of growth and maturation in quite their own way and language is simply one of these structures" (Rieber, 1983: p 2).

Tracing the existence of language to children's early behaviour was initiated by Tomasello (as cited in Goldin-Meadow, 2007). The early behaviour of children to point towards particular objects of interest was regarded as a form of language. Research then focused on children's first vocabulary formation, which was based on the items towards which they commonly pointed. This suggested a link between early learning and the pointing or signalling of children since signals (gestures) leads to word forming in children. Further research indicated that based on the objects at which a child pointed at around age 14 months; one could predict the magnitude of a child's vocabulary when they reached age 42 months (Goldin-Meadow, 2007).

Children's pointing (gesturing) to items is indicative of a thought-process in that they observed the items to be of particular significance to them. These signals could be regarded as primitive forms of communication for children, which would later result in speech and sentence formation. These signals suggest that the behaviour is linguistically derived and assists in extending the child's linguistic ability. This linguistic ability occurs in two ways. One approach is that the signals enable the child to receive verbal feedback from an adult, in the form of naming the object at which they pointed. This advances the child's linguistic skills. The second approach is that the signals have cognitive components, as it provides the child with problem-solving abilities. Both approaches, however, suggest that these signals enable learning in children (Goldin-Meadow, 2007).

When comparing how learning occurs in deaf children's ability to signal, there are several interesting findings. Firstly, deaf children used signalling, which is indicative of linguistic skills, similar to hearing children. The method for which signals are used for deaf and hearing children is analogous in their suggesting of particular items, yet deaf children commonly use combined signals. Hearing children learn from verbal feedback, while deaf children learn to form their own signing language. Signalling can be considered the initial language ability and when the child can communicate verbally, signals become applicable for other purposes. Signals are used as a language and a way of communicating with adults. These signals facilitate the process whereby language is formed, and speech follows. The use of signals can also be considered a means by which the child is aware of social consequences. This also sparks a debate on whether language is formed or is a learnt behaviour. However, the issues raised suggest a combination of both ideas (Goldin-Meadow, 2007).

Research pertaining to language learning and attaining language has remained an important issue. Berk (2004) refers to studies by Cane (1976, as cited in Berk, 2004) in which case studies of children's language attainment were examined. The first study was of Victor, aged between 10 and 13 years who was a modern Tarzan, in that he lived by himself in the woods, without any language skills to communicate with people. After being exposed to language and receiving rigorous instruction, he failed to grasp language skills and only managed to acquire signals. It was speculated that Victor could have had autism, which would have explained why he was not able to learn language effectively. The second study was the case of Massieu, who was deaf but had hearing parents. His family used a home signing system to communicate, which consisted of signalling patterns, as his other siblings were also deaf. At age 13, he received language instruction for several years. He was able to perform basic language skills, yet struggled to use the correct word order correctly. The other case studies consisted of a girl who had been abused and another who was only exposed to language in later years (Berk, 2004).

All these cases were similar in terms of the individual's linguistic shortages, which seemed to be connected to their verb agreement. Their delayed acquisition of language appeared to be negatively influenced by their verb agreement, as they were only able to grasp the basics of language. This suggests that delayed acquisition limits the language being fully learnt, as the more advanced language skills are not as easily acquired in delayed situations. When this was explored with deaf adults who had varied experience with American Sign Language, similar findings were found with regard to verb agreement problems. This allowed the researchers to conclude that for verb agreement to be fully integrated, it needed to be acquired within a specific period. Furthermore, deaf adults who acquired American Sign Language at later ages also experienced problems with language, such as memory and comprehension. The problems associated with verb agreement affect the development of language in both deaf and hearing individuals when their language learning has been delayed. This issue seems to persist despite language instruction (Berk, 2004).

During schooling, children learn about different language structures such as vocabulary and semantics, and this informs their sentencing. This advances their understanding of words and situations in which words are used. The use of syntax, which refers to the way in which words are structured in a sentence, advances as children continue in their schooling. Through these different language structures, comprehension is utilised to infer meanings from words and contexts. Additionally, individuals are guided by the use of pragmatics, which refers to culturally explicit social gatherings, which informs how they communicate. Pragmatics becomes more advanced as individual's progress through life. It should, however, be noted that how children and adults advance cognitively and linguistically is dependent on many factors and thus it is expected that not all individuals will develop similarly (Ormrod, 2008).

As interest grew about whether language influences thought, many theories were developed, but a prominent one was the Sapir-Whorf Hypothesis. This hypothesis was based on the following ideas: Languages differ in terms of its semantic partitioning of the world, the language structures used can influence how one understands and perceives the world, and individuals who speak different languages will perceive the world differently (Gentner & Goldin-Meadow, 2003; Li, 2003). This was the hypothesis by which the research conducted by Gentner and Goldin-Meadow (2003) was informed. In their study of individuals from different language groups, they found that these individuals responded similarly on non-verbal tasks while very differently on verbal tasks. This allowed them to conclude that there is a universal cognition, while semantic structures differ across languages, indicating that the conceptual structure is not universal. This finding implies that semantics and conceptual structures are not dependent on each other (Gentner & Goldin-Meadow, 2003).

In conjunction with this, Boroditsky (2011) argued that thinking is shaped by language, and different languages allow individuals to vary in their cognition. The way people make sense of the world and how they relate to each other differs due to their languages. For example, some languages (such as the Kuuk Thaayorre language spoken in Pormpuraaw) have space and time inherent in their conversations while other languages (such as English) do not use it if it is not necessary. Thus, some languages require individuals to be well acquainted with space and time compared to languages that do not frequently use it when speaking. Moreover, languages are also influenced by memory, as some languages are more detail-orientated than others (Boroditsky, 2011).

There is a link suggested between surroundings and an individual’s level of intelligence. Dickens and Flynn (2001, as cited in Schaie, 2008) suggested this link and indicated that people who had higher levels of intelligence tended to choose surroundings that enhanced their intelligence and operated at a higher pace. Changes in the surroundings of the individual could also lead to additive intelligence, which is referred to as a multiplier effect. This corresponds to what occurs in a social context when groups of individuals have a rise in their intelligence, and the general intelligence of the society consequently grows. This improves the level of their interactions and promotes changes on a societal level across time. This suggests that there is an interchange between ability and surroundings, because of the increase in the intellect of individuals. Dickens & Flynn (2001, as cited in Schaie, 2008) argued that there was a specific element (x factor) within individuals’ surroundings that increased particular behaviours. This element was how he explained the influence of culture and environment on an individual’s behaviour and cognitive development. Moreover, intelligence (both fluid and crystallized) is influenced by culture in early adulthood, as this is the period in which an individual’s intellect is increasing (Schaie, 2008).

When considering the connection between language and cognition, different disciplines are important, namely anthropology, linguistics, and psychology. During the 1950s, questions relating to this connection were explored, and the main concern was whether language was a separate ability from other cognitive abilities or whether it was part of cognitive ability. Within cognitive research, there are two approaches to understanding how the various cognitive components work, namely general purpose and mental modules.

The general purpose approach identifies cognitive abilities as part of an overall system that works together to solve problems (Harris, 2006).

The mental modules approach identifies cognitive abilities as separate entities that work in distinct ways. In both approaches, the view of language is either a distinct mechanism or a general mechanism of the cognitive system. The mental modules approach was influenced by neuropsychology and certain areas of the brain responsible for memory, language, and so forth were regarded as distinct functions. The 1950s were also crucial because of the study of generative linguistics, initiated by Noam Chomsky. He believed that language could be considered to function in a way that is similar to parts of the brain. Chomsky reasoned that children had an innate ability to learn language and had some language awareness. This awareness was related to linguistic concepts such as nouns, verbs, and grammar. Chomskyan linguistics was therefore centred on the idea that language was innate and unique and is divergent from cognition (Giri, 2003; Schaie, 2008).

While behaviourism prevailed during the 1950s as the leading psychological theory, this theory did not support Chomsky's view of language, because he did not accept the notion that children learnt language by imitation. This argument was guided by the proof that children would make use of incorrect phrasing and language use, which they would not have learnt by imitating adults. The incorrect phrasing and language use of children was referred to as linguistic over-regularisation, because the language used by children was influenced by how they heard and perceived adult's language, rather than by mere imitation (Harris, 2006).

Chomsky believed that children had a need to learn language and their ideas about language were not related to other cognitive abilities. Since he believed that children had this distinct ability that was not only related to language, it was labelled the language acquisition device and later, universal grammar. His contribution to linguistics was referred to as generative linguistics because it was based on the idea that the mental models that assist in constructing grammatically correct sentences within language needed to be examined. Additionally, syntax was identified as a significant ability within language, and he argued that it was free from other influences and developed independently. This syntactic ability was therefore referred to as the 'autonomy of syntax' hypothesis (Giri, 2003; Harris, 2006).

The development of the computer was the awakening of a new era, in that cognitive process and systems were compared to those of the computer. This computer metaphor allowed Chomsky and psychologists to describe how language processes function and how they are sub-divided into tasks such as language comprehension. While this new movement in thinking invoked many changes, Chomsky's hypothesis argued for the emphasis of language being learnt and opposed the notion that individuals are born with ability and information. This hypothesis was further influenced by Piaget, who emphasised the idea of language being learnt. He also recognised the relationship between language and cognitive abilities. Piaget argued that language was associated with the cognitive changes that occur when children move from sensorimotor processing to formal and logical processing in adulthood (Harris, 2006).

During the period of the 1970s and 1980s, there were two views present within the language and cognition debate. Psychologists viewed and highlighted the similarities between language and cognitive abilities, while linguistics tended to highlight the unique nature of language. This debate persisted and led to three different developments during the 1980s and 1990s. The first was connectionism in the 1980s, which described the connection between language and cognition using a computer analogy. The computer was regarded as a source of intelligence and had processes that allowed different actions to occur. Connectionism viewed language as learnt and acknowledged a link between language and cognition. The theory of connectionism was identified because it explained how many processing units (such as neurons in the brain) assisted in creating networks that aided parallel processing and intelligent behaviour. This processing system allowed individuals to differentiate between ambiguous words and past tense (Harris, 2006).

The second development was called cognitive linguistics, which arose in the late 1980s and linked language and cognition. Cognitive linguistics was associated with functionalist linguistics, which claimed that language was needed to provide effective communication. Linguistics argued against the validity of syntax, and this initiated the notion that cognitive psychology be combined with linguistics. Thus, cognitive linguistics was labelled as such because there was a need to acknowledge that cognitive abilities were important in how sentences were constructed and understood in linguistics. In the 2000s, cognitive linguistics had prevailed as a dominant field but has however remained outside the realm of linguistics. The period of the 1980s until present represents the modularity of mind, which argued that language was innate and distinctive, and thus it was dissimilar to cognition (Harris, 2006).

The third development was the cognitive neuroscience movement in the 1990s. The cognitive neuroscience debate suggested that language and cognition had a dynamic interaction. This interaction was multifaceted and involved identifying similarities and differences between these two systems. Neuroscientists and neurobiologists explored the notions of the autonomy of syntax and language being innate. They also identified the areas of the brain responsible for language and indicated that it was flexible, suggesting they could change, and this allowed them to argue that language learning was influenced greatly by experiences. They proposed that language was epigenetic because it was argued that an individual’s behaviour advanced in terms of genes and environmental influences in prenatal and postnatal periods. This correlated with Piaget’s theory that biological systems progressed, which caused cognitive abilities to change over time. This was further supported by the finding that the genes in human genomes cannot be responsible for processes such as language proficiency or use, and the area in the brain responsible for language is subject to many exchanges between interior and exterior surroundings (Harris, 2006).

Language originates from an individual's mind and is connected to cognitive processes such as reasoning and perception. A significant study of the association between language and cognition is cognitive linguistics. This form of linguistics is aimed at exploring and understanding the mechanisms of the mind. Cognitive linguistics argues that language cannot merely be reduced to language structures, but it includes the use of cognitive processes. Thus, the use of these language structures is guided by cognitive thought. An additional view of language was demonstrated by iconicity in language, which explored language structures and motivation (Radden, 2008).

Cognitive linguistics was initially shaped by psycholinguist's work on categories and categorisations, which relates to the processing of language. Categories are defined as a grouping of things that are similar in nature. The ability to categorise things into meaningful units is what relates cognition to language. These categories require rational thought and simplify the means by which one refers to groups of things. It should, however, be noted that various languages tend to categorise things differently; thus categories cannot be compared across languages or culture. An example of this categorisation is the grouping of food. In addition to this, taxonomies refer to categories that have attached specific meanings, such as tablecloth. This refers to the combination of the two words that form a whole (Radden, 2008).

The last type of categorisation is prototypes, which refers to the most appropriately suited items that form a category. This links to the way in which the category is understood; for example the category of father is understood as a working man and parent. The ability to use this categorising concept reflects three issues that connect cognition to language. Firstly, words can have social connotations, and thus there should not only be a focus on the linguistic properties. Secondly, there are many categories that are meaningful because of the associated prototype, and this creates a better understanding of what the categories entail. Thirdly, the use of categories is guided by metonymy, which is focused cognitive conceptualising and not linguistic features (Radden, 2008).

The evaluation of conceptual structures involves the following: domains, conceptual frames, scripts, mental spaces, and conceptual blending, all of which form part of cognitive linguistics. In order for categories to exist, these overarching conceptual structures must be considered (Radden, 2008).

Domains refer to either physical or psychological concepts, such as touch in relation to the domain of hand, or faith in terms of the domain of spirituality.

Conceptual frames are a selection of words that share a particular relation to each other. For example, a frame on commerce would be: buying – goods – selling – money (Radden, 2008).

A script refers to sets of behaviour associated with particular events. For example, a script on having supper implies that the following activities occurred: purchasing food items, preparing these food items, cooking, and eating the prepared food. Thus, when one reads that someone has eaten his or her supper at home, the previously identified events are assumed to have occurred.

Mental spaces refer to how individuals present their views. For example, when referring to what particular authors claimed, an argument of a particular viewpoint is expressed.

Conceptual blending refers to words being combined to form a new word, such as brunch, which refers to a meal time situated between breakfast and lunch (Radden, 2008).

Furthermore, the relationship between language and cognition was demonstrated by the use of metaphors and metonymy. Conceptual metonymy and metaphors are identified as figures of speech, and because of their conceptual nature, they are a combination of thought and language. The first is conceptual metonymy, which is similar to a frame or domain, but when one aspect is mentioned, other similar aspects are assumed. For example, when mentioning Aristotle, philosophical writings are implied. The second is metonymy, which refers to the use of a word that has a double meaning, such as “the rifle came late to the parade ground”, which allows for several meanings to be generated for the word rifle. It refers to a weapon, the soldier, and the army frame. The third is conceptual metaphors, which refers to phrases or sentences that are conceptual and are linked to abstract thinking. An example of this would be: “These readings have given me a new perspective and some good ideas”. This reference to the readings is how conceptual metaphors are depicted (Radden, 2008).

Furthermore, Reddy (in Radden, 2008) described communication and language in terms of a conduit metaphor, which essentially involves the forming of ideas into words, moving ideas around, and then selecting ideas from words. This is similar to the way in which verbal reasoning is described, which involves activities related to encoding, decoding, and sending information. These metaphors express the notion of loading, removing, and transmitting information. Radden (2008) argued that this metaphor of language ability and learning is wrong, and he identified reasons why individuals tended to believe these metaphors of language learning and ability. His reasoning was that words are more meaningful in a sentence than in isolation. Since communication of language is possible through sound waves, the use of metaphors allowed for a simplistic approach to understanding a physically complex function. Thus, Radden argues that language ability is inaccurately viewed as merely encoding, decoding, and transmitting information (Radden, 2008).

Cognitive linguistics also includes the concept of cognitive grammar, which incorporates the dimensions of syntax and cognition. Cognitive grammar was influenced by cognitive linguistics, which argued that there was meaning to be implied from the language used and not only from the linguistic properties. An example of cognitive grammar would be the expression of a “glass half full” (Radden, 2008, p. 23). The exploration of cognitive linguistics is vital in comprehending the complex nature of language and cognitive skills. Cognitive linguistics emphasises the many facets in which language and cognition combine to inform meaning as well as the use of reasoning.

An interesting aspect of cognition is the link between language and the perception of emotion. It was found that the use of emotion words, such as “thankful” in a text, can increase an individual’s accuracy because it targets his or her recognition memory. Furthermore, studies have found that the area in the brain that accounts for perceptual features also accounts for the linguistic emotions perceived on faces (Lindquist & Gendron, 2013). The result of such research lies in the confirmation of the link between language and cognitive structures.

These different movements of either confirming or denying the existence of language and cognitive abilities being related and interdependent are necessary. This emphasises the different schools of thought with regard to the language and cognition debate. It does, however, illustrate the various aspects of the debate that are significant.

3.5 Conclusion

This chapter provided an interesting look into the cognitive realm of testing and theorising. The way in which the various theorists constructed their model of intelligence allows one to grasp the complexity involved with categorising the numerous cognitive skills available. It also creates an understanding of the interrelation between various cognitive abilities, such as reasoning and language, which does not occur in isolation.

The exploration of the significant psychological theories that acknowledge the cognitive development of individuals creates a relatable dialogue from which one can introspect into either personal cognitive abilities or observed abilities in others. These theories are important when considering the sample that is used for this study, as it moves across the range of young adult to older adult. The theories related to cognitive development informs the understanding of their performance on measures related to cognition.

The remaining portion of this chapter focused on the relationship between language and cognition. This section connects with the previous chapter because it confirms the link between aspects of language and cognition. This relationship is more pronounced when assessments are involved, and the recognition of cognitively influenced language skills creates for a better understanding of cognitive assessment.

The important characteristic of language relating to cognitive skills is what underlies the factors presumed to be intrinsic in the ECT. Furthermore, the realisation that reasoning can be depicted as language skills is often what separates psychology from disciplines such as linguistics and language studies.

CHAPTER 4: THE RESEARCH RELATING TO THE ENGLISH COMPREHENSION TEST

4.1 Introduction

This chapter explores studies relating to reading comprehension and how this comprehension links to cognitive skills. The chapter advances the discussion of language structures having a cognitive capacity and not merely being features of language.

The discussion then proceeds to the aspect of cognitive skills that are expressed in a verbal format and are used for the assessment of reasoning. This refers to the construct of verbal reasoning, which is very broad and is inclusive of many intellectual tasks of a verbal format. To grapple with this dynamic construct of verbal reasoning, important studies relating to the construct of verbal reasoning ability are explored. This exposition should establish the link between language, cognition, and reasoning.

Since this study hypothesises that the dominant factor present in the ECT is verbal reasoning, this knowledge is instrumental in guiding the discussion of the findings. This will allow a better comprehension of the core construct dominating this study.

This chapter is concluded with a review of the initial findings of the ECT, which provides valuable insights into its dimensions. Based on the discussions of language and cognition, the proposal that verbal reasoning is the prominent construct being measured by the ECT was discussed.

4.2 Reading Comprehension

Since the ECT involves a reading text, it is essential to explore the literature on reading comprehension. This provides insight into an inherent aspect of the test and allows one to establish the link between all possible constructs emerging from a comprehension test.

Language involves many cognitive and linguistic components, yet fundamentally, it is involved in the process of comprehension. Comprehension commonly takes the form of reading comprehension, as this is the method by which an individual gains an understanding of the text. Reading comprehension is defined as a process of reading that allows two critical processes to take place: decoding and comprehension (Kendeou, Van den Broek, Helder & Karlsson, 2014; Pretorius, 2002).

Decoding is defined as “the process whereby the written letters and words are translated into language” (Oberholzer, 2005, p. 21). Consequently, cognitive psychologists have always had an interest in reading comprehension as it involves cognitive skills such as decoding, memory, reasoning, and knowledge (Kendeou et al., 2014; Van den Broek & Gustafson, 1999). Reading comprehension is crucial when assessing individuals with tests that contain reading texts (Kendeou et al., 2014; Pretorius, 2002). When individuals read texts, they comprehend these texts by accessing memory that provides them with meaning and allows them to make inferences from the text (Kendeou et al., 2014; Van den Broek & Gustafson, 1999). This is how comprehension assists in the process of understanding and the assignment of meaning to a text (Kendeou et al., 2014; Pretorius, 2002).

Comprehension is closely associated with the ability to draw inferences from the text. The process of reading involves the reader establishing a causal relationship between the texts and consequently creates coherence within the text. Coherence in this context refers to an individual’s ability to be consistent and logical in how they make meaning of the reading text. This strategy of reading assists readers in memory or recall tasks, as they can remember events that have many causal connections. Thus, individuals will be successful in reading comprehension if the cognitive processes they use during reading allow them to create an image that is coherent and easy to recall for either answering questions or retelling (Kendeou et al., 2014; Van den Broek & Gustafson, 1999).

All of these components do not function independently but are fundamentally linked and function interdependently. Individuals who have cognitive and academic challenges often also have poor verbal skills. This deficiency is often linked to poor reading skills. Reading is classified as crystallized intelligence because it is a learnt behaviour and allows facts and information to be obtained. This aligns with Cattell's definition of crystallized intelligence, and thus any malfunction in reading ability will cause a decline in reasoning skills. Since reading is related to verbal cognitive functioning, a study conducted on two tests, namely the Wechsler Intelligence Scale for Children-Revised (WISC) and the Cognitive Ability Test (CAT), was examined to determine how this factor influences performance. From the examination of these two tests, it was found that social experience was plaguing how individuals performed on the verbal intelligence measurement. These tests were criticised for their reliance on the socially gained knowledge that determined how individuals performed on the test. An inspection of the content of the items suggested that individuals would be able to answer the items correctly if they were exposed to a particular environment (American society). When assessing the individuals' performance on the other cognitive measures; they concluded that poor verbal skills led to poorer performance on the CAT (Langdon, Rosenblatt, & Mellanby, 1998). Studies such as these increase awareness of the instruments used by psychologists for assessing cognitive functioning. This information creates awareness of how performance on verbal reasoning is influenced by the content of the items.

4.3 Comprehension as a Cognitive Exercise

Language comprehension is commonly regarded as the evaluation of linguistic tasks and skills due to the emphasis on language. It was, however, found that language comprehension measures aspects of general cognition (Gernsbacher, 1990). Through the examination of language comprehension, it became evident that there are cognitive tasks that are crucial to individuals' understanding of texts. This consequently led to the development of the structure-building framework. This framework suggests that the aim of comprehension rests in the reader's ability to create a coherent mental representation (Gernsbacher, 1990; Kendeou et al., 2014).

The process of language comprehension is facilitated by memory nodes as they are responsible for initiating the process, and once a foundation of information is formed; all the relevant information is incorporated into this structure. Only when other information is obtained, either irrelevant or not connected to the existing information, a substructure is formed that stores this information. It is towards the completion of the comprehension that the structure created would resemble a tree with many branches (substructures), in that all the information received was stored according to its relevance (Gernsbacher, 1990).

Two key activities of comprehension are suppression and enhancement. Suppression involves the reader's ability to identify irrelevant information within the text, while enhancement involves the reader's ability to identify relevant information within the text. These two activities assist the reader in forming an accurate representation of the text he or she has read. Individuals who are less skilled at comprehension are often characterised by their inability to suppress irrelevant information within the text, resulting in an inaccurate representation of the text read (Gernsbacher, 1990; Kendeou et al., 2014).

Reading involves an individual creating meaning from the text using either restricted background knowledge (referred to as a text base) or volumes of background knowledge (referred to as a situation model). The representation formed by the reader will, therefore, be positioned between the text base and the situation model, depending on the reader's knowledge, the text, reading goal, and his or her motivation (Kendeou et al., 2014; Van den Broek & Gustafson, 1999). This representation is how meaning is created by individuals and the means by which they understand the text they have read.

Koch (2015) reflected on issues experienced during her longitudinal study on additive bilingual education in South Africa, which required research on the English and Xhosa versions of the Woodcock Munoz Language Survey. The issues that were raised related to fundamental aspects of South African Education such as the South African Educational testing programme, also referred to as the Annual National Assessment, and the South African Department of Basic Education's Language in Education policy. These two educational factors (Annual National Assessment and the Language in Education policy) severely impacted the research and the validation process of the South African adapted Xhosa and English versions of the Woodcock Munoz Language Survey (WMLS). These predicaments included delays in the data collection process and complications with the adaptation of the English and Xhosa versions of the WMLS because the language of education had changed during the period of the study (Koch, 2015). These educational factors have impacted many school-going individuals and as a result, need to be considered in the South African context. This, therefore, emphasises the need to validate cognitive psychometric assessments in South Africa, paying close attention to factors such as education that may either hinder or promote achievement in psychometric assessments.

4.4 Verbal Reasoning

Historically, the theory of reasoning was conceptualised in philosophical writings, and the nature and way in which individuals reason was coupled to processes such as logic. Reasoning is explained as the process by which inferences are made from the information provided. The process of logical reasoning is executed by using either deductive or inductive reasoning, which are the two methods for reasoning logically. Deductive reasoning is when the individual must make inferences from the information given and naturally leads to the conclusion being drawn. Inductive reasoning, however, requires the individual to infer from the information provided but the conclusion that is formed is more speculation than guaranteed. The difference between the two is, however, small as studies have found strong relations between these two methods of reasoning (Lohman & Lakin, 2009).

Two important psychological theories have guided how psychologists view reasoning, namely the mental rules and mental models theory. The mental rules theory considered reasoning to be comprised of the following processes: encoding information into working memory, obtaining conclusions from the information using abstract thinking, and accessing working memory for any inconsistencies. This theory is linked to the process of deductive reasoning and most errors, according to this theory, are due to working memory incapacity (Lohman & Lakin, 2009; Manktelow & Chung, 2004). The mental models theory (Johnson-Laird, 2004) is also associated with the process of deductive reasoning. This model operates as follows: the premise must first be assembled. The premise is then used to create a model. While constructing a conclusion that explains the relationship but does not reveal this relationship in the premise, different models are then identified, which could challenge the existing model or conclusion. The models are then compared for similarities. Mistakes made in reasoning, according to this theory, are due to working memory because multiple models need to be created to ensure that the conclusion reached is valid (Lohman & Lakin, 2009; Manktelow & Chung, 2004).

It is imperative to note the level of cognitive awareness of individuals when they are reaching certain conclusions. These levels of cognitive awareness are captured by utilising either tacit or intentional reasoning processes. Tacit and intentional reasoning processes operate on different levels of consciousness. Tacit processes operate unconsciously, while intentional (explicit) processes operate consciously. Tacit processes usually occur in instances where automatic responses are given, while intentional processes require thought and careful action (Lohman & Lakin, 2009). These processes are observed in different situations requiring individuals to reason, and these levels of consciousness often support their response.

A system by which reasoning tasks can be achieved requires processes such as selective encoding, selective comparison, and strategic combination. Selective encoding is a process by which an individual can separate relevant and irrelevant information. Selective comparison, on the other hand, requires individuals to access their long-term memory and compare the relevant information to the problem. This comparison is based on the conclusions made about the relationship between concepts and identifying the best possible solution to the stated situation. Strategic combination refers to the use of deductive reasoning in tasks that require individuals to derive a conclusion (Lohman & Lakin, 2009).

According to Stenberg (1986), inductive reasoning errors are associated with problems in selective encoding and comparison, while deductive reasoning problems refer to strategic combination. Within both these forms of reasoning, there is an underlying mechanism by which this reasoning occurs. This mechanism is known as working memory and literature has identified clear links between working memory and reasoning ability, specifically that of fluid reasoning. Thus, it is essential to acknowledge the influence that working memory can play when interpreting tasks based on reasoning ability. An integral component of reasoning is verbal reasoning, which is commonly measured in tasks related to sentences, comprehension, and vocabulary (Lohman & Lakin, 2009).

Analogies have commonly been used to measure verbal reasoning. The use of analogies in intelligence testing was based on identifying relationships between items, where the analogy would signify the relation between said items. The use of analogies was important when measuring intelligence and contributed to psychometric assessments. This use of analogies can be classified as inductive reasoning, and was categorised as fluid intelligence when researched. It was, however, noted that some analogies would require crystallized intelligence (Holyoak, 2012). Additionally, verbal analogies assessments were linked to processes such as decoding and comprehension. The format of these analogies is usually depicted in a stem using two or three words. The answering format requires the individual to complete the omitted words in this stem. For example: "Hen is to chick as mother is to ?" This, therefore, necessitates the use of verbal ability, predominately in the form of vocabulary, and general intelligence as relationships between items need to be interpreted (Roomaney & Koch, 2013).

Verbal analogies assessment is connected to verbal reasoning and has been criticised for drawing on crystallized intelligence. This presents particular challenges as exposure to different sources of information can cause discrimination among individuals, particularly those of different cultural groups. On the positive side, however, verbal analogies can be instrumental in measuring verbal ability and intelligence if observed to be unbiased. In light of this, the structure associated with verbal analogies is usually multidimensional as it is connected to both verbal ability and general intelligence. Vocabulary hinders the use of verbal ability and thus the need to use words that are not too challenging is encouraged (Roomaney & Koch, 2013).

Koch (2015) identified specific challenges associated with the verbal analogies scale for the Xhosa version of the Woodcock Munoz Language Survey in South Africa; primarily that these analogies in English presented problems. The analogy type of questioning could not be done for the Xhosa version as it was not understandable or a familiar way of speaking within this language phrasing. For this reason, the Xhosa verbal analogies scale preserved an analogy type of questioning but was completely different to the English verbal analogies test with regards to the method in which the questions were phrased. Based on this, the Xhosa version was regarded as easier due to the clues given in the items (Koch, 2015). This emphasises the difficulty associated with an analogy type of assessment for verbal reasoning within South Africa, as it is not a commonly used way of comprehending information.

Research on the verbal analogies scale of the Woodcock Munoz Language Survey for both English and Xhosa versions in South Africa indicated that there were DIF items found, and equivalence could not be established across these two test versions. It was, however, discussed that this was a common finding for verbal analogies scales, especially across different language versions as the meaning of items would not necessarily be the same across languages. Moreover, verbal analogies assessment is regarded as a biased form of testing verbal reasoning. This was emphasised by the finding that even when the DIF items were removed from the scale, structural equivalence across these two language versions could not be established. The explanation for this was that verbal analogies were not easily translated across English and Xhosa. They were, therefore, not measuring the same concepts (Roomaney & Koch, 2013).

In a study exploring gender bias for an intelligence measure (the Slovak Version of Intelligence Structure Test 2000 revised), the review of literature indicated that males performed better than females on verbal assessments such as verbal reasoning. In their study, however, the findings regarding the verbal analogies subtest were imperative as this subtest had the highest amount of DIF items when compared to the other subtests of the intelligence measure. It was also found that the females outperformed the males, and on inspection of the items causing the DIF, it was found that the content was related to female-preferred activities (Kohut, Halama, Dockal, & Zitny, 2016). This reinforces the concern that the verbal analogies test can be problematic, even when assessing gender performance.

Due to the psychometric issues that confront South African test developers, some developers embarked upon the challenge to address these issues and created South African specific instruments, such as De Beer (2004, 2011) and Bekwa (2016). Even in 2016 (Bekwa, 2016), the need to address cultural biases in cognitive assessment in South Africa through the use of psychometric assessment is still essential. Bekwa‘s (2016) study was unique in that she created African items (using African art) to assess non-verbal figural reasoning. This study echoes a new era in test development, and is one in which the ECT wishes to unite. Moreover, the use of the Rasch model in Bekwa's doctoral dissertation was stressed as vital for the South African context and assisted in the development of the instrument. Bekwa's study emphasises the need for test development in South Africa that addresses the multicultural context (Bekwa, 2016). It should, however, be noted that although verbal assessment can be problematic in contexts such as South Africa, it is an essential part of testing cognitive ability and cannot be completely replaced by non-verbal assessment (Lakin, 2012).

From these extracts of literature on the concept of verbal reasoning, it becomes evident that the notion of verbal reasoning is not explicit. The definition is derived from elements of reasoning and authors using this concept tend to merely identify it as verbal ability. The difficulty with defining the construct of verbal reasoning is that it encompasses many elements of intellectual functioning. There are numerous studies that make use of the term verbal reasoning and many cognitive tests are labelled as verbal reasoning (Differential Aptitude Test, Senior Aptitude Test,), yet many resemble verbal ability. Most verbal reasoning tests that are developed in the UK or USA make use of verbal analogies, which is not a common assessment method in educational settings such as schools. This study hypothesises that it will measure this construct of verbal reasoning, yet not in the traditional sense of analogies. For this reason, this study will define verbal reasoning as the composition of deductive and inductive reasoning skills to identify plausible responses to verbal stimuli.

4.5 Studies on Verbal Ability

Internationally, there are mixed feelings regarding brain injury research and the effect that minor brain injury (such as loss of consciousness for a few moments) could have on an individual's normal functioning. Interestingly, the USA reports a relatively large number of minor brain injury cases, in which the individuals do not usually require much treatment. Since this area of research is ambiguous, a study was conducted to determine the effect that minor brain injury has on an individual's verbal ability (Clark, Garner, & Brown, 1992).

This study intentionally focused on an individual's ability to complete verbal analogical reasoning tasks. The reason for choosing these kinds of tasks was based on the combination of verbal ability and abstract reasoning. Verbal ability is considered unperturbed by minor brain injury while abstracting is severely affected by any brain trauma. Additionally, verbal analogical reasoning tasks correlate with intelligence, which provides insights into intellectual functioning (Clark et al., 1992).

The verbal analogical reasoning tasks were evaluated according to the individual's ability to encode, infer, map, apply, compare, justify, and respond to information. The findings obtained in this study provided constructive feedback for cognitive research and testing. The results indicated that verbal analogical reasoning was affected by minor brain injury. When comparing the performance of the affected and unaffected individuals, the affected individuals made more errors and were slower on tasks. The specific aspects of concern for the affected individuals were their ability to encode, infer, and compare (Clark et al., 1992).

This finding is significant for the ECT as the proposed construct of verbal reasoning involves the cognitive processes of encoding and inferring. Since it is not known whether individuals had any minor brain injury, it must be accepted that it could be a threat to the validity of the instrument. It is important to acknowledge that this could influence how individuals complete the ECT as it requires these reasoning skills that might have been impaired if they suffered any minor brain injury.

A study on a verbal reasoning test established that there was a strong correlation between individuals' performance on the pre-and post-testing (Strand, 2004). This indicated that performance on verbal reasoning tasks did not change over time; it remained relatively stable. Regarding comparing individual's performance based on their schooling, it was found that high school individuals performed better, while primary school individuals performed poorer. Since individuals are still acquiring reasoning skills when they are in primary school, they are expected to perform poorer, while high school individuals have developed more reasoning skills. These findings demonstrate that reasoning skills are only stable in older individuals (Strand, 2004). This understanding of verbal reasoning has informed the concept of the construct to be evaluated in the ECT, and since the sample involves adults, it is expected that their verbal reasoning ability will have stabilised.

In a study on cognitive functioning, 15-year-olds were asked to complete several cognitive assessments. Their performance on the various cognitive assessments suggested that their ability to reason abstractly developed before their verbal reasoning ability. This was based on the low scores these individuals obtained on the verbal component of the reasoning test (VESPAR) (Langdon, Rosenblatt, & Mellanby, 1998). Since the age group of the participants that completed the ECT is above 15 years old, it can be implied that both these abilities should have been developed. This finding is, however, significant when considering the concept of verbal reasoning and its cognitive development.

A study on whether demographic factors affect the cognitive abilities of Latvian people was explored. It was found that factors such as region, age, and gender had an effect. The age pattern confirmed previous findings of an increase in cognitive ability for people aged 20 to 40, while people between 41 and 97 showed a drop in cognitive ability. This also indicated that crystallized intelligence and fluid intelligence increase and decrease according to certain age groups. In the verbal ability test, there was, however, no differences observed for ages, and this lack of decline suggests that it is an aspect of crystallized intelligence (Griskevica & Rascevska, 2009). This confirms the findings of the previous study that adults have a stable verbal reasoning ability.

A study on the verbal reasoning subtest of the DAT-M provided interesting results regarding the construct validity of this subtest. The development of the verbal reasoning test involved items that were deliberately created to be uncomplicated so that the focus was not on the difficulty of the items, but on the concept of reasoning. When evaluated, this test had high reliability coefficients. In terms of the construct, the test assessed abstract thinking with the use of verbal items, in the form of analogy type questions. The verbal reasoning subtest also indicated strong relations with intelligence measures. When comparing this subtest to the WISC, it was found that there was a relation to the verbal comprehension factor (this was comprised of vocabulary, comprehension, and similarities) as well as general IQ (Cooperman, 1980). This finding supports other research that acknowledges the relationship between intelligence and verbal reasoning.

When exploring the literature surrounding cognitive testing, various facets could possibly influence an individual’s performance. The context of South Africa is important, specifically because the test was designed for this population. Although, HIV/AIDS is not part of the present study, it is worthwhile to consider some research conducted on this disease, as it could be an extraneous variable. It should also be noted that an individual’s HIV/AIDS status is private, and obtaining disclosure permission and authorisation can be both problematic and challenging. This knowledge would also be considered discriminatory (Tomu, 2013).

The actuality of the extent of the spread of HIV/AIDS in South Africa is well known and in 2008, South Africa was recognised as the country with the highest HIV/AIDS population (Mupawase & Broom, 2010). Several studies have confirmed that HIV is responsible for causing cognitive impairment in affected individuals. One of these cognitive impairments is a form of dementia associated with this cognitive decline, and many complex neurological tests have been designed to examine this. Additionally, HIV/AIDS affects the area of the brain that is responsible for language. This initiated the investigation by Mupawase and Broom (2010) to explore the impact of HIV on language and communication skills within South Africa. Since time and resources are common issues that plague the assessment of HIV/AIDS individuals, the use of the Cognitive Linguistic Quick Test (CLQT) was introduced as it could reduce testing time considerably (Mupawase & Broom, 2010).

The CLQT was tested on HIV-infected South Africans to assess whether it would be a viable screening test for cognitive impairment, particularly as it is not a locally developed assessment measure. The results of the study indicated that the CLQT could be used as a screening measure and that it was able to identify a decrease in the affected individual's memory, executive functioning, and psychomotor skills. Language skills were, however, not impaired, and this allowed them to conclude that language for everyday functioning was not affected, while language involving inductive reasoning would be problematic (Mupawase & Broom, 2010).

This is a rather interesting finding, and when considering that the ECT could be a possible measure of verbal reasoning, it can be considered a potential threat to validity. This is, however, a threat to any cognitive test in South Africa, and there are no foreseeable means by which this can be eliminated since an individual's HIV status is private and irrelevant to employment decisions (as one may not discriminate based on HIV status). This assessment concern adds value to the discussion on cognitive testing and could not be overlooked due to its controversial nature.

The Senior South African Individual Scale-Revised (SSAIS-R) was standardised and showed evidence of a relationship between verbal ability and academic performance. The verbal scale is comprised of vocabulary, comprehension, similarities, number problems, and story memory, which encompasses verbal intelligence. The results indicated that there were correlations between the verbal scale and the following subjects: language, geography, mathematics, general science, and accounting. This verbal scale also accounted for the academic performance of Grade 9 students (Marais, 2007).

In Marais' (2007) study, the regression analysis of the DAT-S and the SSAIS-R had several interesting findings. It was found that the vocabulary and verbal reasoning subtests of the DAT-S contributed to the verbal intelligence scale. Furthermore, the reading comprehension subtest of the DAT-S contributed to the non-verbal intelligence scale. Moreover, the total score of the SSAIS-R, the vocabulary, verbal reasoning, and reading comprehension subtests assisted in predicting general intelligence in learners. This indicated that the SSAIS-R assisted in determining verbal and non-verbal intelligence. The verbal reasoning and vocabulary subtests also assisted in predicting performance in the subject English. In addition to this, the verbal reasoning subtest correlated with the Natural Science, while the subject Human and Social Sciences correlated with the vocabulary and verbal reasoning subtests (Marais, 2007).

The research on the tests of the DAT form L indicated that the verbal reasoning test was the best predictor for assessing academic performance and success. It was, however, only limited by an individual's lack of language skills (Owen, 2000). This corresponded to research by Lakin (2012) in which it was also found that verbal reasoning was linked to academic success. The influence of language on verbal reasoning tasks is significant and must be acknowledged as an important variable within the ECT.

A validation study was conducted on the Undergraduate Medical and Health Sciences Admissions Test (UMAT), which is used in Australia and New Zealand. The results indicated that the logical reasoning tests had strong correlations with the verbal and numeric ability tests, had only slight correlations with the abstract reasoning test, and had no relation to the emotional intelligence test. Additionally, it was found that the verbal ability test was also the only construct within the selection battery that was able to predict a section on understanding people (Griffen, Carless, & Wilson, 2013). This study highlights the impact of reasoning, specifically verbal reasoning, in selection batteries for specialist careers. It shows the flexible nature of this construct in that it can be used in various environments, and emphasises the imperative nature of this ability, as it can be a deciding factor for professional careers.

A study on the perception of intelligence used Chinese and English students to rate items that they believed to be related to intelligence from different tests. The factor analyses conducted on their responses identified two major factors, namely verbal and non-verbal reasoning skills. An interesting finding was that the English group rated verbal reasoning to be more relevant to intelligence than the Chinese group, and this difference in opinion was attributed to the difference in schooling and mental effort by these groups of students (Chen & Chen, 1988). This reiterates the importance of verbal reasoning.

4.6 The Construction of the English Comprehension Test

The ECT was created in 2010 by research psychologist D.E. Arendse, who is the author of this thesis. The reason for developing the ECT was prompted by the observation that many candidates participating in selections were unsuccessful due to difficulties in comprehending the English in the cognitive tests they completed. The ECT was empirically designed as a basic assessment of English language and comprehension skills. The items were therefore created by the researcher to correspond to the comprehension piece and language skills to be assessed. All of the items in the ECT are multiple-choice questions, with the exception of four written responses. The scoring method of these multiple-choice items is dichotomous (which scores one option as correct and all other distracters as incorrect). The origin and intended use of the ECT was therefore to serve as a screening tool which could locate possible English comprehension problems. This has remained the intended use for the ECT, except that it may best serve as a possible screening tool for Verbal Reasoning.

The ECT version 1.2 was the initial test version, and several changes were made after analyses were conducted. The administration remained the same, but several changes were implemented. Firstly, the number of instructions on the test booklet was increased. Secondly, the comprehension and language sections of the test (ECT 1.3) were demarcated more clearly to avoid the confusion witnessed with the ECT version 1.2. Thirdly, the answer sheet was replaced by a scanner-friendly version as a means of shortening the time spent on data capturing. Fifthly, the problematic items that were identified were edited and a few items were added. This increased the total number of items from 39 to 42 in ECT version 1.3. Sixthly, the time limit of 45 minutes was removed as performance on the test could be attributed to functioning under pressure, which would affect the validity and reliability of the test. In ECT version 1.3, the time period in which participants complete the test was documented to assess the average time required to complete the ECT (Arendse & Maree, 2017).

The data collection for the ECT version 1.2 transpired in 2010 while the ECT version 1.3 transpired in 2011 during various selections. The ECT version 1.2 and 1.3 were, however, exposed to the same selection batteries. The tests that were used in the different selections were the: Academic Aptitudes Tests (AAT) 1: non-verbal, AAT 2: verbal reasoning, AAT 3:

vocabulary, AAT 4: reading comprehension, AAT 5: numeric comprehension, DAT 2: verbal reasoning, DAT 3: non-verbal reasoning-figures, DAT 9: mechanical insight, DAT 10: memory, SAT 2: calculations, SAT 4: comparison, SAT 5: pattern completion, SAT 6: figure series, SAT 7: spatial 2D, SAT 8: spatial 3D, and SAT 10: short-term memory.

The exploratory factor analyses conducted on the ECT version 1.2 and 1.3 initiated the proposed study on the validity and reliability of the ECT. The results identified a prominent factor within both these versions. Initially, the researcher identified this construct as semantics (language construct) because the items that loaded on this factor related to meaning and paragraph comprehension (Berk, 2006).

Further investigation and reading (Gernsbacher, 1990; Polk & Newell, 1995) caused the researcher to hypothesise that this factor was not semantics, but verbal reasoning (the psychological construct). This conjecture was influenced by the fact that verbal reasoning is a process whereby verbal images are transformed into semantic images (Polk & Newell, 1995). This essentially means that individuals deductively solve problems while using linguistic tools (Polk & Newell, 1995).

4.7 The Initial Findings of the English Comprehension Test

The initial analyses conducted on the ECT version 1.2 and 1.3 revealed interesting results. Since the ECT version 1.3 had no time limit, the time it took the last person of the different groups to complete the ECT was recorded. The results indicated that it took an average of 74 minutes for individuals to complete the test. The shortest time recorded was 55 minutes and the longest time was 113 minutes. The implication of this finding was that the previous time limit of 45 minutes cannot be considered for the next version of the ECT as it will be biasing individuals. For this reason, no time limit of any kind will be applied but the recording of test times will be continued to establish a possible future time limit (Arendse & Maree, 2017).

The structure of the data for the ECT version 1.2 and 1.3 was examined by performing exploratory factor analyses. The process of factor analyses for the ECT version 1.2 and 1.3 involved two methods of extraction: principal components analysis (PCA) and principal axis factoring (PAF), which were compared and assessed regarding their assessment of factors and item loadings. PCA was chosen because it is different to factor analysis in that it usually assumes no relationship exists (Field, 2009). PCA is, however, suggested when there is no prior knowledge of the scales being explored. PAF is regarded as a method of exploratory factor analysis and can be used when the data are not normally distributed (Yong & Pearce, 2013).

Since the development of the items was done empirically, there was only speculation that they should be measuring English comprehension. Thus, both methods were used. This would aid the development of the ECT. The process of factor analysis for the ECT version 1.2 and 1.3 was the same for both the PCA and PAF. The method of rotation used for both forms of factor analysis was the promax rotation method, because this method of rotation produces correlated factors, which is essentially what the researcher intends to observe. The appropriate matrix for this study was the pattern matrix, because it is used for interpretation when oblique rotation has been used (Costello & Osborne, 2005; Hair, Black, Babin, & Anderson, 2009).

The exploratory analyses conducted on both the ECT version 1.2 and 1.3 revealed noteworthy findings. Three factors were retained for the ECT version 1.2 that explained 25% of the variance for both PCA and PAF. In the pattern matrix for the PCA, 27 items loaded as follows: 18 items on factor 1, five items on factor 2, and four items on factor 3. The pattern matrix of the PAF had 17 items that loaded as follows: 5 items on factor 1, eight items on factor 2, and four items on factor 3. The PCA pattern matrix had more item loadings and had a very strong factor 1, while in the PAF analysis there was a strong factor 2. Despite these differences in factor loadings, there was clear evidence of unidimensionality, with a strong factor emerging from both analyses (Arendse & Maree, 2017).

There were three factors retained for the ECT version 1.3 that explained 24% of the variance in the test for both the PCA and PAF. The pattern matrix of the PCA had 25 items that loaded as follows: 13 items on factor 1, five items on factor 2, and seven items on factor 3. There were cross-loadings of item 7 on factors 1 and 3, and item 8 on factors 1 and 4. The pattern matrix of the PAF had 20 items that loaded as follows: five items on factor 1, seven items on factor 2, and eight items on factor 3. The PCA and PAF analyses revealed that both had a dominant factor, which emerged as factor 1 in PCA and factor 3 in PAF. This was indicative of unidimensionality as there was clear confirmation of a dominant factor in both analyses (Arendse & Maree, 2017).

The PCA and PAF analyses for both versions contained fairly similar results, yet the PCA pattern matrix had more item loadings for both versions. The correlation matrix for the PCA and PAF for both versions had a similar trend in that some of the factors of the PCA analyses were less related, while all the factors in the PAF analyses were related (Arendse & Maree, 2017).

The MDS PROXSCAL solution was used because it seeks to find the best fitting model using the smallest number of possible dimensions. The fewer the number of dimensions there are, the easier it is to read and interpret the data (Busing, Commandeur, & Heiser, 1997).

PROXSCAL is a program available on SPSS version 16.0 and offers users better graphs in their output (Zhang & Takane, 2010). MDS requires the data be specified as metric or non-metric. The data that were used for this study were metric data and the measurement level used was interval (Jaworska & Chupetlovska-Anastasova, 2009; Moroke, 2014).

The data were used to create proximities, and only one matrix was used as there was only one data source. The Euclidian model was used to calculate the dissimilarities of the data in dimensional space using the distance model (Arce, De Francisco, & Arce, 2010; Moroke, 2014). The scree-plot and stress function were used to determine the acceptable dimensions for the data (Jaworska & Chupetlovska-Anastasova, 2009; Moroke, 2014). For the ECT 1.2, the S-stress measure indicated that the variance accounted for by the three dimensions was 98% (Tucker's coefficient of congruence), and four dimensions accounted for 99%. Additionally, the Dispersion Accounted For (DAF) for both the 3- (96%) and 4- (97%) dimension solutions were excellent, as this value should be close to 1 (Arce et al., 2010). The 4-dimension solution was chosen for the ECT 1.2 because the error in the distance model is reduced to 1% (variance that is not accounted for in model) and it is an acceptable fit for both Stress-1 and S-stress measures (Arendse & Maree, 2017).

For the ECT 1.3, the S-stress indicated that the variance explained by both the 4 and 5 dimensions was 99%, using the Tucker's coefficient of congruence. The DAF for the 4 and 5 dimensions was 98%, which was an excellent value (closer to 1). Since both the stress functions indicated that a 4-dimension solution was a good model fit and the variance explained by this model was 99%, the 4-dimensional solution was chosen for the ECT 1.3. The MDS graphical output identified that the proximities of the data for the ECT version 1.3 were better than the ECT version 1.2. This suggested that the data improved from the one version to the next (Arendse & Maree, 2017).

The labelled factors which are based on the loadings of the EFA (PAF) for the ECT version 1.2 were as follows: factor 1: Vocabulary, factor 2: Reasoning and factor 3: Deduction. The labelled factors of the EFA loadings (PAF) for the ECT version 1.3 were as follows: factor 1: Vocabulary, factor 2: Deduction and factor 3: Reasoning. The similarities between the factor structures for both versions endorsed the presence of definite structures within the ECT, despite the few changes between the versions (Arendse & Maree, 2017). The evidence of a dominant factor was identified in all the analyses conducted for both ECT versions 1.2 and 1.3. This was a critical finding as it assists in the development of the ECT for future usage. This dominant factor was initially speculated to be a language construct, but it was not clear what it could be. Since it was an exploratory study, the results needed to be investigated more thoroughly and more analyses needed to be conducted to provide substantial information on the constructs being measured by the ECT (Arendse & Maree, 2017).

The initial findings on the ECT suggested that the main factor could be verbal reasoning, because of the elements of which the ECT is. This notion was informed by the understanding that language acquisition and academic success is attributed to the verbal reasoning skills required by individuals to complete comprehension tasks (Lakin, 2012). In light of this, verbal reasoning is identified as a level of intelligence, which is significant. The most stressed argument of the verbal-reasoning hypothesis is that the cognitive processes involved in deduction, such as encoding and re-encoding, are the same as those occurring in language comprehension. This, therefore, follows that deductive reasoning is executed by using linguistic skills (which also includes visual-spatial skills); nevertheless, the underlying processes involved are cognitive in nature (Polk & Newell, 1995).

Given this, there is a ladder formation that commences with intelligence at the top, descending into reasoning, and then downwards into verbal reasoning. If an analogy were to be used, intelligence would be the tree, and one of its branches would be reasoning. One of the leaves that make up this branch is verbal reasoning. It can be construed that the leaves (verbal reasoning) cannot exist without the branch (reasoning), and hence the branch cannot exist without the tree (intelligence).

4.8 Conclusion

This chapter placed the study and the construct of interest, verbal reasoning, into perspective. The investigation of literature regarding reading comprehension allows one to identify the inherent reasoning situated within reading materials and tests that use reading texts. The exploration into how comprehension is not a linguistic task but is associated with reasoning skills, is what promotes the hypothesis of the ECT being linked to reasoning.

The structure-building framework demonstrated the connection between cognition and comprehension. This paved the way for the construct of verbal reasoning, which can be regarded as a combination of verbal skills linked to both inductive and deductive reasoning. The case studies discussed in this chapter focused on studies investigating verbal constructs which had conducted similar validation analyses.

This verbal reasoning construct is hypothesised to be the main factor in the ECT, which was demonstrated in the review of the exploratory factor analysis conducted. This study invited further investigation into the construct validity of the ECT as well as determining the factors being measured by the ECT.

This chapter provided insight into the background of ECT and places this study into context. This background information provides insight into the selection of methods used as well as the motivation behind this doctoral study. Additionally, the review of this literature serves to identify possible confirmation of other factors innate to the ECT.

CHAPTER 5: MESSICK’S UNIFIED THEORY OF VALIDITY

5.1 Introduction

The way in which validity has been understood and defined has undergone many changes over the past 20 years. The way in which tests has been described and explored in terms of their validity has evolved dramatically. This implies that validity cannot be determined from a single source as it does not suffice and validity information requires a process of several forms of information. This comprehensive form of information is more substantial than a single source (Yun & Ulrich, 2002).

Validity can be understood in various ways depending on the research question at hand and theoretical underpinnings of the study. Since the context is crucial to understanding validity, the choice of validity theorists is based on the aim of the study. In this study, construct validity was envisaged, and consequently Messick was identified as instrumental to framing the study.

The theoretical framework of this study was, therefore, Messick’s unified theory of validity (Baghaei, 2008; Messick, 1995, 1996; Ravand & Firoozi, 2016) because it was the best suited theory to guide the understanding of the findings. The choice of using Messick’s framework was influenced by the concept of validity and the need for a comprehensive method in which to explore test development and use. To further substantiate the use of Messick’s theory, the themes identified in his framework link to the issues manifesting from language testing as well as testing of cognitive abilities or skills.

Validity is part of measurement theory. The generalisation of validity studies is limited because it only applies to the population it was tested on and the reason or aim for the testing. From Messick’s viewpoint on validity, the notion of validity is not merely left to a validity coefficient but allows one to dwell on the issues emerging from using the validity information. This extension of validity to explore factors relating to inferences and conclusions based on the results of validation is what separates Messick's theory from other measurement and validity theories. In a society consumed by inequalities and injustice, the need to explore the consequences of validation results is crucial. It informs decision-making and informs what conclusions can be drawn from a test (Kane, 1992, 2006; Messick, 1995, 1996; Yun & Ulrich, 2002).

These theoretical underpinnings of Messick's unified theory of validity are only some of the many reasons behind the use of this framework for this study. Other validity theories fall short of considering the contextual issues that are a significant aspect of this study, specifically because this test was piloted in a multicultural and multilingual country (South Africa). When considering the construct validity of the ECT, contextual issues are not only applicable to the analyses done on the instrument but also pertinent to the construct itself. Additionally, there is a great difference in terms of using Messick's unified theory of construct validity compared to other validity theories as this theory allows for a comprehensive look at important features that create an integrated view of the construct being measured. The contribution that will be made by exploring the results through the systematic lenses of Messick's unified theory of construct validity will enable the ECT to be fully explored and dissected in terms of whether there has been sufficient evidence gathered to suggest that the ECT is valid.

This chapter explores the concept of validity, Messick's theory, and challenges regarding Messick's theory. This provides more insight into the reason behind the selection of Messick's theory as the theoretical framework underpinning this study.

5.2 Defining Validity

Historically, the emphasis of ensuring that tests were performing as they should was largely guided by the development of tests in order to approximate succinctly measures of intended constructs. To assess whether a test was measuring the intended construct effectively, the construct of a correlation coefficient was established with the associated formula. This correlation coefficient was influenced by Pearson and assisted in both the development and evaluation of tests. The correlation coefficient therefore allowed researchers to assess whether the constructs being measured by the tests were related to similar or the same constructs established by an existing test. This influenced the notion of validity in that "a test is valid for anything with which it correlates" (Guilford, 1946, p. 429. in Wolming & Wikstrom, 2010). This new-found analysis of validity was, however, short lived as it provided a very narrow interpretation of the validity of the test. This notion caused validity to be considered a property of the test. Thus displaying any evidence of a relationship with test scores and evaluating it against external criteria was deemed as sufficient evidence for test validation (Baghaei & Amrahi, 2011; Wolming & Wikstrom, 2010). The need for a more substantial interpretation of validity was brewing, which led to the American Psychological Association to develop a preliminary guide by which researchers could evaluate and assess tests to ensure their validity and reliability (Hamavandy & Kiany, 2014; Kane, 1992).

Previously, four forms of validity were stressed, namely content, predictive, concurrent, and construct validity (Smith, 2001; Wolming & Wikstrom, 2010). Since predictive and concurrent validity employed external measures, they were later combined to form criterion-related validity. The three forms of validity were regarded as the "holy trinity" (Wolming & Wikstrom, 2010). This simplistic view of validity then shifted to a broader view, which identified different types of validity, namely content validity, criterion-related validity, and construct validity. Content validity referred to the performance of individuals in a specific domain, criterion-related validity referred to the tests being able to predict future achievement, and construct validity referred to tests that could make inferences about cognitive abilities (Baghaei & Amrahi, 2011; Kane, 1992; Smith & Smith, 2004; Morrow, 1981; Wolming & Wikstrom, 2010).

The error in reasoning during this period was the notion that validity could be divided into separate forms, each responsible for particular functions. This meant that one form of validity was sufficient to declare a test valid. This caused problems for test users as many issues started arising from this method of deducing validity. Construct validity then became the most significant form of validity, and literature pertaining to this grew as the prominence of construct validity shifted to a higher platform (Shepard, 1993). Within this period of the 20th century, Alfred Binet was recognised as influential in his creation of intelligence tests, especially because of its focus on construct validity (Hamavandy & Kiany, 2014).

This movement was followed by Cronbach and Meehl, who endeavoured to provide a more comprehensive definition of validity. Cronbach and Meehl (1955, p. 257) held the position that tests were not validated, but rather validity served as a ‘principle for making inferences’, which essentially referred to the uses and interpretations made by tests and their scores. This ultimately led to consequential validity, as these tests had consequences that needed to be examined. Cronbach and Meehl promoted the use of construct validity over criterion and content validity. This caused a change in how validity was interpreted, and the meaning then moved from being a property of a test to the interpretation of test scores (Embretson, 1983; Kane, 2006; Messick, 1995; Wolming & Wikstrom, 2010).

In addition to this, Cronbach and Meehl coined the term ‘nomological net’ (meaning lawful), suggesting that the theorised method of thinking about the construct involved considering various relationships that could impact how the construct was explored. They theorised a method in which the construct was hypothesised. Firstly, the relationships assisted in outlining the construct and aided in its existence. Secondly, variables needed to have a predictive relationship between the construct, which formed part of construct validation. Thirdly, construct validity was comprised of current, inter-item, and inter-test correlations. These correlations needed to be high or low depending on the conjectured relationship between the variables and the construct. Their shortcomings were, however, that they oversimplified the process and this approach did not resonate well with other researchers, and caused a shift in thinking (Embretson, 1983; Wolming & Wikstrom, 2010).

A unified definition of validity was then introduced by Messick (Messick, 1995, 1996; Smith, 2001). Messick's theory of validity was instrumental in guiding and validating language tests, specifically because of his focus on social consequences (Van der Walt & Steyn, 2008). In this framework, construct validity was interpreted in terms of two layers, namely interior and exterior layers. The interior layer referred to the construct being investigated, while the exterior layer referred to the linking of the construct to other similar assessments. The structural and substantive factors which were later labelled as "construct representation" and "nomothetic span", were addressed within the internal component of the test. The construct representation and nomothetic span essentially involved ascertaining whether the construct of the test was the intended construct and whether it was accurately embodied in the test, both internally and externally (Embretson, 1983; Shepard, 1993).

Within this framework, the validation of the construct consisted of four facets. The first facet involved defining the construct, which referred to the explanation of what the test instrument was measuring. The second facet involved preparing a hypothesis which would examine the construct, and informed how the construct needed to be examined. The third facet involved the use of statistics to explore the construct in a scientific way, as different methods provided different sources of information. The final facet involved obtaining sources of information to sustain the construct and the inferences that were made about the construct (Yun & Ulrich, 2002).

Construct validity became promoted as the fundamental element of validity. Construct validity has been theoretically argued as the most important aspect of validity and was defined as the ability of a test to measure that which it is intended to measure. Content and criterion validity were identified as part of the process of construct validity. This promulgated the idea that conclusions made on the basis of test scores were to be taken seriously and needed to be examined carefully. These ideas resonated with society and as a result, test development and use acquired much attention. The evaluation of test scores and the conclusions drawn from these scores were stressed as the need to avoid negative societal implications of tests was emphasized. Thus, the more tests were used in any context, the more questions were raised about the test quality, its fairness, and whether it was used appropriately (Baghaei & Amrahi, 2011; Kane, 2006; Messick, 1995; Palmer & Bachman, 1981; Sireci, 2007).

When considering the history of how validity was conceptualized, it becomes apparent that the evaluation of tests is imperative. This evaluation of tests is often referred to as test validation and involves an examination of the properties of the test, by exploring the validity and reliability. Validity relates to the extent to which the test is accurately fulfilling its intended purpose while reliability refers to the test’s ability to consistently produce the same results (Martin & Savage-McGlynn, 2013). Validity is comprised of many elements, such as face validity, content validity, construct validity, and predictive validity. Validity is also a principle that language tests must adhere to but this process is not as simplistic. The validation of language tests are carefully considered because of both legal and cost implications. The test needs to consequently be examined and deemed valid before it can be used (Van der Walt & Steyn, 2008; Weir, 2005).

There are two key problems which affect validity, namely construct underrepresentation and construct-irrelevant variance. Construct underrepresentation refers to a test that is unable to measure the construct effectively. The implication of such a test is that it will incorrectly measure the abilities of an individual and the conclusions made by the test will not be a true reflection of the theorised construct. Construct-irrelevant variance refers to factors not originally part of the construct that affect the measurement of the construct. It can either cause individuals to perform better (indicating that they experience the test as easy) or they can perform poorly (individuals experience the test as difficult) (Baghaei, 2008; Hamavandy & Kiany, 2014; Messick, 1995; Weir, 2005). These problems related to validity need to be considered as they influence the interpretations made by the test.

Reliability is essentially evaluated through a statistical procedure, and it is commonly used as one of the principles by which language tests are evaluated. A test measures a certain construct and a score is derived from the test based on how well or badly the individual performed. The score is used as a means of reducing subjectivity in testing, and it is then used in the process of validation (Weir, 2005). Weir labelled reliability and its related elements as “scoring validity” (Weir, 2005, p. 16). Reliability therefore evaluates the consistency of this test score.

Construct validation, in the modern day, therefore concerns the inner and outer aspects of the test to be assessed. Evidence needs to be provided on both aspects because this will allow for convergent (the existence of the same construct in another measure) and discriminant (the absence of the construct in an opposite measure) validity to be established (Embretson, 1983; Messick, 1995; Shepard, 1993).

It should, therefore, be understood that a test cannot be perfect and all elements explored will not always yield positive results (Davies, 2003). For this reason, validity is regarded as a process and not an immediate or complete answer. The process of test validation, therefore, informs the use of tests and indicates the elements that need to be re-evaluated to improve the test (Davies, 2003; Kane, 2006).

In the historical quest of establishing validity, the concept of validity grew and became the most influential concept in psychometrics. The definitions concerning validity have shifted through the years and have moved from broad to more specific concepts to limit confusion. The importance of establishing validity was emphasised with its increased use in educational settings, and awareness was raised about the possible implications that could arise from the use of tests. These implications spread to a societal level, and the need to ensure that the tests being used allowed for correct inferences to be made became more prevalent. The processes of validity and test validation were identified as the means by which correct inferences could be made (Sireci, 2007).

In an article, Sireci (2007, p. 477) referred to validity as “the ultimate challenge for a psychometrician”. This challenge resonates with the work of many others (such as Messick, 1995, 1996; and Kane, 2006) in that validity revolves around obtaining evidence that will endorse the use of a certain test for a specific purpose. The evidence, therefore, supports the purpose of the test instead of trying to establish if a test is completely valid. Sireci (2007) identified four issues relating to validity; firstly that validity is related to the purpose of a test and is not an inherent property of the test. Secondly, several types of evidence should be obtained to establish the use and appropriateness of the test. Thirdly, the purpose of a test should be defended by the evidence obtained (this evidence should be able to justify the use of the test for the purpose), and fourthly, validity is an ongoing process and cannot be limited to a single statistical output (Erguven, 2014; Sireci, 2007).

Based on the historical context of validity and the shift in thinking, the need to evaluate the social implications of tests has been amplified, especially with all the cross-cultural findings suggesting an adverse impact has occurred for minority groups. For this reason, Messick's theory of validity was chosen as the theoretical framework of the study. This theory is comprehensive and will facilitate a beneficial discussion on the ECT.

5.3 Messick's Theory of Validity

When measuring psychological constructs, they require one to operationalize the construct that needs to be assessed. The operationalizing of these psychological constructs enables there measurement and easier to reference than when one merely speaks of their existence. Messick's progressive matrix identified sources of justification and the two core functions of testing, test usage and interpretation. The sources of justification involved evaluating the evidence and the consequences related to testing. The results of testing are evaluated by exploring the interpretations made by the test results and its usage. Within test interpretation, construct validity and value implications are considered. Within test usage, construct validity, the relevance, and utility of the test are explored as well as the social consequences of its use (Van der Walt & Steyn, 2008).

Messick's theory of validity is described as a unified view of the validity processes and involves the incorporation of many methods to provide empirical evidence of validity. This theory furthermore emphasises the value attributed to the interpretation of the evidence obtained (Hamavandy & Kiany, 2014; Messick, 1995; Weir, 2005).

	Test Interpretation	Test Use
Evidential Basis	Construct Validity	Construct Validity & Relevance/Utility
Consequential Basis	Value Implications	Social Consequences

Figure 1: Messick’s (1995) Facets of Validity Framework

In Figure 1, the different aspects of Messick’s theory of validity are displayed. Messick explained test validation as a theoretical and practical means by which validity can be addressed. He identified four regions that can be used to establish validity: construct validity, the relevance of utility, value implications, and social consequences. These regions are shown in the figure above, and act as a guideline that psychologists can use when addressing test development issues. The core of his theory rests on his view of construct validity (Hamavandy & Kiany, 2014; Messick, 1995).

The fact that construct validity is indicated in two of the blocks is not so much redundant, but rather implies that more evidence regarding the construct needs to be obtained. The block identified under test use, situated across from evidential basis is what stimulated much research regarding techniques for construct validity. This block expects that the developer ensures that the construct being measured is both relevant for its intended purposes and will be used in a fair way. The utility refers to the use of the test for a defined purpose, which must be justified and evidence of this must be demonstrated. This means that sophisticated techniques should be used as proof of the construct validity of the test to be utilised in a particular context. This is where instances of bias in testing and within the test itself should be inspected. After the analysis of the test's usage and relevance, the consequences for society must be reflected on. This means avoiding tests that could cause adverse impact on individuals being tested (Messick, 1995; Shepard, 1993).

A critique of demarcating the different elements into blocks is that researchers or developers might avoid completing the requirements of all the blocks and merely attend to those they regard as necessary. The issue of allocating construct validity in two blocks concerning test interpretation and test use for evidential basis can be problematic. This being that researchers might disregard the additional construct validity as already obtained, which is not what Messick intended, as its duplicate status implies that additional findings should be obtained. Messick's framework was labelled a progressive matrix suggesting that the researcher or developer should move systematically from one block to the next, thereby advancing the information obtained every time. All the blocks are therefore interdependent and rely on each other to create a comprehensive picture of the test being analysed (Shepard, 1993).

Thus, all aspects are important and considering only the scientific aspects such as construct validity information and neglecting the value and social consequences will distort the meaning of the interpretations being made as well as the accuracy of how the test is being used. The fact that validity continues throughout also gives rise to the fact that there might be more questions than answers and construct evaluation is necessary, as different theories may need to be tested. As a result, the first block can be used when a test is being researched and explained, but when the test needs to be used to select individuals for either employment or education purposes, the other blocks need to be considered, as their inferences became visible (Shepard, 1993).

Although there are demarcated blocks in the matrix, the boundaries are relatively fluid. The value implications refer to the method in which scores are interpreted. The scores should be used to assess the construct or ability accurately, thereby giving an accurate interpretation of what the scores mean. This said, the evaluation of any issue that could affect an individual's test score must be considered. Since Messick's theory influenced language testing and the validation of language instruments, the emphasis of social implications is of particular interest when validating language instruments (Van der Walt & Steyn, 2008).

In Messick's framework, test administration and issues related to the environment of the testing are considered. This incorporates both the ethics and the possible effects of testing, such as teaching following a testing outcome. The issues relating to the individual completing the test are also considered, such as any physical or emotional issues, for example as their ability to concentrate during testing, and prior testing exposure (Weir, 2005).

Messick's theory is centred on three values, namely the benefit or lack of harm that the tests should have towards individuals, the useful nature of the test, and the removal of bias, by advocating fairness. This theory, however, failed to provide guidelines by which these values could be practically attained. His recognition of the social implications of tests, which are often referred to as consequential validity, has added great value to the testing community and initiated the formation of two significant theories of validity, that is Weir's socio-cognitive model and Kane's argument-based model of validation (Hamavandy & Kiany, 2014).

Messick's theory achieved two purposes. Firstly, construct validity became elevated to be the most important concept of validity as had now been described as all encompassing. Secondly, his theory transformed validity thinking to consider the implications of scores, which includes social consequences, utility, and values. The repercussion of not considering the values attached to particular tests was the reason why Messick included this aspect under test interpretation and consequential basis. This requires test users or developers to consider values concerning the use of a particular test in certain populations for either employment or education (Shepard, 1993).

Messick identified six aspects of construct validity, namely content, substantive, structural, generalisability, external, and consequential facets. These aspects were identified to attend to construct validity in a way that allows one to address the unified notion of construct validity (Baghaei & Amrahi, 2011; Messick, 1995, 1996; Ravand & Firoozi, 2016; Smith, 2001; Smith & Smith, 2004).

The content facet of construct validity entails the various aspects of the test content that should be appropriate for the construct being measured. This content includes being applicable, having a descriptive nature, and may include the methodological value of items. The judgement made on the basis of the content should be done by experts in the field. The methodological value of items includes items that should be edited to eliminate ambiguity and address the inappropriate level of items for certain persons. This also involves the technical aspects of the items in the test. The test developers need to be wary of content that threatens construct validity, such as construct underrepresentation and construct-irrelevant information. This, therefore, implies that there should be a sufficient spread of items in terms of their ability to measure the construct at different levels of ability (Baghaei, 2008; Baghaei & Amrahi, 2011; Messick, 1995, 1996; Ravand & Firoozi, 2016; Smith, 2001).

The substantive facet of construct validity refers to the confirmation of the construct used in the test. Methods need to be employed to ensure that the person completing the test is engaging with the content and not with other issues. This will ensure that the construct to be measured is being assessed. This method usually takes the form of multiple-choice questions, and distractor analysis is then used to provide confirmation that the distractor items either succeeded or failed in diverting the person from the answer. This also refers to the method in which the person completed the assessment, to establish whether his or her responses relate to the difficulty of the items. The substantive facet, therefore, refers to the processes by which individuals employ in completing the assessment, are aligned to the construct being investigated (Baghaei & Amrahi, 2011; Messick, 1995, 1996; Ravand & Firoozi, 2016; Smith, 2001).

The structural facet of construct validity refers to the way in which the test is scored. The reasons for scoring should be informed by the structure of the test. This implies that if the test is made up of several constructs, then these constructs should each be scored separately and not as a total score. Only unidimensional tests may, therefore, have a total score, as there is only one construct under investigation. The factors emerging from the test consequently need to be assessed to ensure that the structural aspect of the test is aligned with these factors (Baghaei & Amrahi, 2011; Messick, 1995, 1996; Ravand & Firoozi, 2016; Smith, 2001).

The generalisability facet of construct validity refers to the scores' implications and the deductions made from the results of the test to be able to be generalised further to include aspects of the wider construct that are not specifically tested in the test. This would inform one, by means of the test, if the full extent of the construct can be generalised. This extends the use of the test scores obtained for the person, meaning that one can generalise his or her ability on the construct in a broader sense and not only on the aspects of the construct examined by the test. This, therefore, implies that the performance of individuals should be based on their performance on the construct being measured, which is indicated by the total score. If their performance is due to other factors, this would indicate invariance, and the total score cannot be generalised across settings or persons (Baghaei & Amrahi, 2011; Messick, 1995, 1996; Ravand & Firoozi, 2016; Smith, 2001).

The external facet of construct validity refers to the test's ability to connect with other tests and behaviours similar to the construct being measured. This ability confirms the meaning of the construct across the test and other tests. This aspect of external validity also refers to the test’s ability to separate individuals based on their levels and knowledge of the construct being measured, which would include those who display high, low, or fluctuating levels of the construct. One of the ways in which the external facet of construct validity can be assessed is through Multi-Trait Multi-Method analyses (Baghaei & Amrahi, 2011; Messick, 1995, 1996; Ravand & Firoozi, 2016; Smith, 2001).

The consequential facet of construct validity refers to the intentional and unintentional aspects of testing such as biases that may influence how people perform on the test. This includes the consequences associated with testing for individuals completing the test. This facet promotes the values of fairness, limiting bias, and outlining the authentic and predictive implications the test may have. These aspects may inhibit the validity of the instrument, especially when items used in the test may preference a certain gender, race, or language group (Baghaei & Amrahi, 2011; Messick, 1995, 1996; Ravand & Firoozi, 2016; Smith, 2001).

These six facets of construct validity are instrumental in assessing whether a test has produced sufficient evidence to satisfy the unified notion of construct validity. By assessing these six facets through the use of various statistical analyses, one may argue that a test has attained construct validity.

Although Messick’s theory is theoretically excellent, it does, however, have limitations. Firstly, Messick provided a guide by which validity can be established, yet not a specific way in which to achieve this. Secondly, the issue relating to validity being guided by interpretation has caused many debates (Hamavandy & Kiany, 2014). These challenges are nonetheless important for advancing the validity argument, as it creates opportunities for new methods of theorising.

5.4 Challenges of Messick’s Theory of Validity

Since Messick’s model was very theoretical, it only allowed the concept of validity to be conceptually well understood, while leaving some room for improvement in the empirical working out. This was mostly based on its lack of practical applications to the theoretical roles. Kane (1992, 2006) subsequently added this practical dimension to Messick’s theory and formed argument-based validity. This model rested on two arguments. Firstly, the interpretive argument referred to how tests scores would be used and interpreted, and which decisions would be made based on this. Secondly, the validity argument referred to the measurement process, and argued that its interpretation should be both acceptable and logical. These elements allowed for a practical means by which to evaluate tests for validation. Kane’s theory only falls short of emphasising the consequences relating to testing. This is why Messick’s theory is more relevant in this study; the social consequences of testing regarding language or cognitive instruments in a multicultural and multilingual country cannot be underestimated (Messick, 1995; Wolming & Wikstrom, 2010).

The implications of any test developed require the consequences to be considered carefully and the developer to be cautious about both the use and promotion of the test. Evaluating consequences are therefore an emphasised aspect of test validation in countries such as South Africa. The use of both theories is, however, a better consideration, as Kane’s model was developed on the theoretical underpinnings of Messick’s theory. His practical guidance allows for methodological assistance while referring to Messick’s theory for a more in-depth interpretation of the construct as a whole (Wolming & Wikstrom, 2010).

5.5 Conclusion

This chapter provided insight into the historical context of validity and how the need to define this concept arose. It also emphasised the issues that were identified in earlier periods, but could not be solved due to a lack of guidance.

The uncertainty regarding validity was what ushered in the era of Messick's theory, which grappled with the issues burdening professionals at the time. The fundamental issues of validity such as test use and test interpretation were carefully discussed, covering the essential elements. The emphasis on construct validity and the issues pertaining to ensuring that this form of validity is achieved were explored and attend to one of the aims of the study.

The significant and unique aspect of Messick's theory is the consideration of the social consequences of testing, specifically related to test use and test interpretation. These issues are important for this study, since the ECT could be used for decision making. The recognition of testing consequences is what compels one to consider the possible effects associated with using the ECT in future.

Although Messick's theory is comprehensive in covering elements relating to validity, there are criticisms directed at his theory. These challenges are largely based on his theory being theoretically sound while lacking practical application. These challenges have briefly been explored as a means of comprehensively addressing validity and test validation.

CHAPTER 6: RESEARCH DESIGN AND METHODOLOGY

6.1 Introduction

This chapter outlines the research design of this study and the methodology that was used. The relevant ethical considerations and ethical authorities involved in this study are mentioned, as this is essential to any study involving human subjects.

The various statistical techniques such as Rasch analyses, multi-trait multi-method (MTMM), confirmatory factor analysis (CFA), differential test functioning (DTF), and reliability analysis were conducted to assess the items and constructs present in the test. These statistical techniques were deliberately chosen for the data analysis because they are the means by which the data can be effectively analysed and interpreted as part of validating the test.

These techniques are not essentially related to each other, but they offer valuable insights into the test functioning and are able to provide a comprehensive evaluation of the test. Rasch analyses and DTF are underpinned by the same measurement theory and are focused on the items of the test and their functioning as a whole and across groups. The CFA and MTMM are underpinned by the same measurement theory in that the focus is on the construct and providing evidence of its structure and its relation to similar constructs measured by other instruments. The reliability statistic informs one about how consistently the test is able to measure the specific construct. Although all these statistical techniques speak to different elements and are underpinned by different measurement theories, they are able to inform the development of the test. The information provided on the items and construct relate to the construct validity of the test while the reliability provides evidence of the constructs consistency. Furthermore, the statistical techniques allow for critical validity checking which is crucial when thoroughly examining a test. This would serve to prevent incorrect inferences being made from test outcomes and aligns with Messick's theory of construct validity.

This chapter will explain how these different statistical techniques were interpreted to ensure that the results section is easily understood.

6.2 Primary Objectives of the Research

The aim of the study was to explore the construct validity and reliability of the ECT, which necessitates specific objectives. These objectives will be explained in greater detail in the data analysis section. The specific objectives of the study were to:

1. statistically explore the unidimensionality of items using Rasch analyses;
2. confirm the dimensionality of the scales using CFA;
3. support evidence of construct validity by conducting a MTMM analysis;
4. explore the measurement invariance using DTF; and
5. evaluate the internal consistency of the ECT by conducting a reliability analysis.

6.3 Research Design

The researcher used secondary data analysis (SDA), which is the re-analysis of data to answer an original research question with improved statistical techniques, or answering innovative questions with the use of old data (Glass, 1976). The design of the research involved using all the data obtained for 2010 and 2011 (these were the years in which the data was collected). This allowed for richer information regarding the two years of data collection periods as well as to assess the development of the test through these two test versions.

The design of this study resembles psychometric validation theory, which requires one to perform certain statistical techniques to ensure that the construct and the items are sufficiently examined to make valuable deductions.

6.4 Sampling and Procedure

When the data was collected, the following sampling and data collection procedures were followed. The sampling used was convenience sampling; the candidates in the study were all attending selections (these were various selections conducted at the Military Psychological Institute), thus making them accessible for the piloting of the ECT.

All candidates were therefore either Grade 12 learners or had already completed Grade 12. The data collection for the ECT version 1.2 occurred in 2010 and for the ECT version 1.3 in 2011.

The administration for these test sessions involved test orientation and assisting individuals with the completion of the biographical section of the answer sheet. The ECT version 1.2 had a time limit of 45 minutes imposed, whereas the ECT version 1.3 did not.

The consent for individuals to participate in the research was done before the selection commenced. They were informed of the inclusion of the ECT and were asked to consent for research purposes. After individuals had completed all the cognitive tests in the selection battery, they took a lunch break and thereafter completed the ECT. The reasoning behind this was that the research should not affect their performance on the tests that will be used for recommendation purposes. The issue of fatigue was considered, especially since the ECT could be measuring verbal reasoning.

The study sample consisted of 597 individuals (in 2010) and 882 individuals (in 2011), both males and females. Their age groups ranged from 18 to 52 years old (in 2010) and 18 to 42 years old (in 2011). These individuals were civilians and military members and resided in all nine provinces. All 11 languages were present in the sample for the ECT version 1.2 and 1.3. Since the ECT has two versions that were administered differently (specifically with respect to the time limit imposed), the samples were explored separately and not combined in this study. The sample size required for the CFA to be effective is 100 – 200 cases for 2 – 4 factors (Loehlin, 1997). Thus, the sample size intended for the CFA was adequate.

There are three crucial considerations concerning the use of a convenience sample. Firstly, the findings cannot be generalised as there is a restriction of range. Secondly, the individuals who participated were part of a selection process and thus their motivation levels would be different from those not participating in selections. Thirdly, the researcher was confined to the data that were previously gathered and any additional information needed for the analyses limited the findings (Boslaugh, 2007). These considerations regarding the use of SDA for this study are acknowledged. Due to the nature of the analyses chosen for this study, the data were well within range to be used and did not hinder the analysis techniques.

6.5 Data Collection Instruments

The ECT is an individual test that assesses an individual’s English language ability and comprehension skills. The ECT contains a comprehension section that is comprised of multiple-choice questions. The language section contains multiple-choice questions and a written answer section. This test has been used on individuals from different linguistic and cultural backgrounds in South Africa. Since it is still in development, it has only been used at the Military Psychological Institute (MPI) in South Africa during a few selections for research purposes. The age range that the ECT had been piloted on ranged from 18 to 52 years. The ECT version 1.2 (2010) had 39 items and a time limit of 45 minutes was imposed. The ECT version 1.3 (2011) had 42 items and no time limit was imposed.

The psychometric properties of the ECT have been explored in this study. The tests that were used in this study to establish convergent and discriminant validity are namely: The AAT (non-verbal, verbal reasoning, vocabulary, reading comprehension, and numeric comprehension), Differential Aptitude Tests (DAT) (verbal reasoning; non-verbal reasoning: figures; mechanical insight; and memory) and Senior Aptitude Tests (SAT) (calculations, comparisons, pattern completion, figure series, spatial 2D, spatial 3D, and short-term memory).

The Human Sciences Research Council (HSRC) was responsible for the development of the SAT (Fouche & Verwey, 1982), the AAT University version (Owen & De Beer, 1997), and the DAT Form L (Owen, 2000) tests for a South African population. These tests were designed to assess individuals on different cognitive structures and are used for educational placement and in occupational selections.

Since this study has used several other tests validated within South Africa, it is important to review their reliability. The reliability coefficients for the above-mentioned tests are as follows:

1. AAT 1 = 0. 87

2. AAT 2 = 0. 76

3. AAT 3 = 0. 85

4. AAT 4 = 0. 81

5. AAT 5 = 0. 94

6. DAT 2 = 0. 47 - 0.80 (0.55)

7. DAT 3 = 0. 60 - 0.75 (0.71)

8. DAT 9 = 0. 31 – 0.81 (0.54) (boys)

9. = 0.02 – 0.64 (0.22) (girls)

10. DAT 10 = 0. 75 – 0.84 (0.77)

11. SAT 2 = 0.921

12. SAT 4 = 0. 762

13. SAT 5 = 0. 834

14. SAT 6 = 0. 852

15. SAT 7 = 0. 918

16. SAT 8 = 0. 838

17. SAT 10 = 0. 762

Based on these reliabilities, the more reliable tests, in terms of appropriateness for selections and high-risk decisions, are the AAT 5, SAT 7, and SAT 2. The tests that are not sufficiently reliable for high-risk decisions, but are sufficient for ability tests are the AAT 1, AAT 3, AAT 4, SAT 5, SAT 6, and SAT 8. The tests that are not sufficient for ability tests, but are appropriate for research purposes are the AAT 2, SAT 4, and SAT 10 (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009).

The DAT L tests have reliabilities that are calculated using a range for Grades 10 to 12. The reliability values in brackets are the median reliability coefficients. Based on these ranges, it is difficult to determine the accurate reliability coefficient. The only one of concern is the DAT 9, which has separate reliability values for males and females. The test publishers of the DAT L have, however, informed the researcher that this test is being researched and new reliability values may be established which are updated and not linked to gender (Personal communication with Mind Musiq, 2015).

These tests have been reviewed and classified by the Heath Professions Council of South Africa (HPCSA). These tests are also on their list of classified tests, which was previously the new law (Employment Equity Amendment Act 55 of 2013, section 8) pertaining to test usage by psychologists (restricted use) in South Africa. According to these legislations and ethical guidelines, these tests are then currently on par and provide a suitable basis for comparison (Tomu, 2013). The only consideration regarding these South African tests is that they are relatively dated, and this was considered when comparing these tests to the ECT.

6.6 Ethical Considerations

Ethics is the discipline that endorses good practice in research with human subjects. Compliance to ethical principles is enforced through the submission of a proposal to ethics committees that examine the study design and data collection strategies. This is done to ensure that no harm is incurred by the human subjects involved in the study.

Psychological studies are strictly governed by ethics and psychologists are required to register with the HPCSA as a means of complying with national and international adherence to correct behaviour in research and work with human subjects. The primary researcher is registered with the HPCSA and thus complies with these ethical guidelines (Tomu, 2013).

It is important to note that the correct ethical procedures were followed during the data collection process. When the secondary data were collected, the following ethical procedures were followed:

Confidentiality: The information gathered contained sensitive and identifying information about the candidates that bears no relevance to the study and thus has not and will not be used in any way and will be kept private.

Informed consent: Before data collection commenced, the researcher obtained informed consent from the candidates during their selection process, and thus they were assured that their information would be confidential.

Safeguarding information: Since the data obtained contain identifying information about candidates, the information has and will be protected and used only by professionals.

This study obtained the following clearances necessary to conduct the research:

Permission from the MPI and The Director of Psychology: Since the data were obtained during the selection processes held at MPI, the researcher obtained clearance to use the data for research from both the Institution and the head of Psychology in the Defence Force.

Permission was obtained from Defence Intelligence and Counter Intelligence to assure the Department of Defence that the information published will not be a security risk and does not implicate the Department of Defence in any way.

Clearance was also obtained from the 1 Military Hospital (SAMHS) Research Ethics Committee to assure the Department of Defence that the research conducted was ethical and did not seek to harm any individual that participated in this study.

Clearance was obtained from the University of Pretoria Ethics Committee to ensure that this PhD study meets their ethical requirements. The committee requires that data storage take place at the University of Pretoria for 15 years.

6.7 Data Analysis

The data analysis for this study consists of descriptive statistics, Rasch analyses, CFA, MTMM analyses, DTF, and reliability analyses. These statistical techniques correspond with the objectives of the study. These statistical techniques are discussed separately below in terms of their relevance, theoretical foundation, and advantages. The procedure that was used to conduct the analysis, as well as the way in which the data were reported for each of the analyses techniques used, is indicated within this section.

6.7.1 Descriptive Statistics of the ECT

The ECT contains a biographical section to describe the sample that was tested for each version. This information is limited to basic descriptions of the sample, while all identifying and confidential information is excluded and does not form part of the analyses.

The descriptive statistics were generated on SPSS version 23.0 to describe the sample as follows: total sample size, age, gender, race, home language, and provinces. The ECT is a multiple-choice test, and all the answer options are coded as dichotomous variables. The means, medians, and standard deviations were evaluated for the ECT version 1.2 and 1.3. The normality of data was evaluated by exploring statistics such as Kolmogorov-Smirnov and Shapiro-Wilk analyses for normality (Field, 2005), skewness, and kurtosis.

According to Hambleton and Jones (1993), item response theory does not require a normal distribution because it assumes that high ability individuals will do better on difficult items than less able individuals. Though normality was violated, the use of Rasch analyses is still appropriate for the data analyses to continue.

The CFA was conducted using SPSS version 23 with the IBM Analysis of Moment Structures (AMOS) software (Arbuckle, 2010). This software allows for effective CFA to be conducted and deals with missing data by using the maximum likelihood estimation method of analysis (Little & Rubin, 1987). This method allows all available data to be analysed.

6.7.1.1 The descriptive analysis procedure

The sample was investigated by conducting frequency analyses for each of the biographical items, such as gender, race, provinces, ages, language groups, home languages, first language at school, and second language at school. These tables were used to indicate the distribution and demonstrate how representative the sample of the ECT was for both versions. This allowed one to ensure that all races, languages, and genders were included, allowing for a comprehensive sample. The specific exploration of languages, which were divided into home, first school language, and additional school language, is important for understanding the findings of the different analyses. The candidates' different language backgrounds had a particular influence on their performance on the ECT and assisted in identifying if the test is appropriate across diverse language groups.

The ECT was then analysed in terms of the distribution of the data. The means, medians, and minimum and maximum values were assessed. These values provide a limited insight into the performance on the test. This was followed by assessing the normality of the data for both versions of the ECT; by use of Kolmogorov-Smirnov and Shapiro-Wilk analyses of normality and the skewness and kurtosis were explored. This allowed for a comprehensive description of the data.

The MTMM analysis requires the use of several psychometric tests, as indicated in an earlier section (6.5) of this chapter. The MTMM analysis requires a description of the psychometric tests used in this study. These psychometric tests were analysed in terms of their respective sample sizes and means. These analyses provide an adequate description of the psychometric tests used, as they were only used as part of a comparative study with the ECT.

6.7.1.2 The reporting of the descriptive statistical analyses

To assist in the interpretation of the descriptive statistical analysis results, the results were reported as follows:

1. The Description of The Sample
2. The Gender Distribution
3. The Racial Distribution
4. The Provincial Distribution
5. The Distribution of Ages
6. The Distribution of Language Groups
7. The Distribution of Home Language
8. The Distribution of First Language At School
9. The Distribution of An Additional Language At School
10. The Distribution of Data
11. Tests for Normality
12. Description of Psychometric Tests

These results were presented separately for the two test versions.

6.7.2 Statistical Techniques

The statistical techniques that will be discussed in this section are as follows: the Rasch model, CFA, MTMM analyses, DTF, and reliability analyses. This chapter discusses the theories underpinning these analyses, the procedures used to perform these analyses, and the order in which these findings were reported.

6.7.2.1 The Rasch model

Unidimensionality used to be defined as a single variable, but this definition was broadened to include psychological processes that allow items to function as a single latent trait, meaning that the items reflect some unity. The need to establish unidimensionality in the items of a test is imperative based on various arguments. One such argument is that when wanting to order persons based on the test, it is essential to establish unidimensionality because persons need to be comparable on the same variable. Another argument posed is that it is important to establish unidimensionality when utilising the Rasch model to determine a person's ability on a test because it could lead to biased results for persons and items (Smith, 2002; Smith et al., 2003). In Rasch analysis, unidimensionality is measured with the use of tests of fit (model fit). This allows one to assess whether the items and persons of the model are unidimensional through the fit statistics (Bond & Fox, 2007). Hence, the consideration of unidimensionality is crucial when contemplating the use of the Rasch model to analyse the items of a test.

The core principle of the Rasch model is that an individual's performance on a dichotomous item is the function of item difficulty and person ability. The probability of an individual correctly answering an item is determined by the discrepancy between item difficulty and person ability (Long, Bansilal & Debba, 2014; Long, 2011; Wright, 1997). The Rasch model (Rasch, 1960) was selected as the preferred method of analysis to explore the unidimensionality of items because of two critical reasons. Firstly, the assumption of unidimensionality in the Rasch model supports the theory of construct validity, which is the aim of this study. Secondly, the test information function (TIF) and test standard error statistics in the Rasch model assist test developers to decide more confidently which items should be retained and which items should be removed (Streiner, 2010). These are, however, not the only reasons for using the Rasch model. This technique provides more substantial information about the items and persons than techniques such as item analysis in classical test theory (CTT).

6.7.2.1.1 The theoretical foundations of the Rasch model

In order to understand the reason for utilising IRT, specifically the Rasch model, the theoretical foundation of this theory needs to be explored. In IRT and the Rasch model, the term “theta” is used to describe the underlying trait (Fraley, Waller, & Brennan, 2000; McBride, 2001). Theta has a mean of 0 and a standard deviation of 1 (Fraley et al., 2000). Theta is the term used to refer to the trait after the IRT or a Rasch analysis has been conducted. IRT is based on the principle that the performance of an individual on a test item is based on the level of ability held by the individual and is also related to the probability of the individual correctly answering the items measuring the latent ability. The Rasch model, on the other hand, generates logit measurements from item parameter estimates and person ability estimates in order to compare them on a common scale. IRT and the Rasch model are very similar, yet they differ conceptually in their approach when assessing the items and persons of a test. Firstly, the Rasch model creates item-person maps in which the item difficulty and person ability can be compared on a common scale. IRT does not create item-person maps. Secondly, the Rasch model makes use of a one-parameter logistic model, while IRT has a variation of one to three parameter logistic models. Thirdly, the Rasch model focuses on the data fitting the model. If the data has a poor fit to the model, then the items which do not fit can be removed and this will improve the model fit of the data. IRT however focuses on fitting the model to the data and thus parameters can be added or removed to assist with the model fitting the data. Fourthly, IRT and the Rasch model use different estimation algorithms when analysing the data. This includes item discrimination and guessing, as this is handled differently by IRT and the Rasch model (Acton, 2003; Erguven, 2014; Hambleton, 1989; Linacre, 2012a; Smith & Smith, 2004; Yu, 2013). These are not the only differences between approaches, but they are the most important reasons for preferring the use of the Rasch model over IRT.

As with every analysis technique, the Rasch model has its own assumptions, namely unidimensionality and local independence. Unidimensionality refers to a scale measuring one trait, while local independence is when the probability of endorsing one item is not related to another item (Embretson, 1983; Pae, 2011; Pae, Greenberg, & Morris, 2012; Smith, 2001; Streiner, 2010).

Unidimensionality implies that the items are measuring one latent trait. If unidimensionality is satisfied, then local independence is also satisfied. The acknowledgement of local independence does not, however, imply that the assumption of unidimensionality is satisfied. Local independence is when item answers do not have any relation to each other. This means that items are independent of each other, so there is no relation between any of the answers to the items. When questions or items in the test rely on one another to be answered correctly, then the assumption of local independence has been violated (Erguven, 2014; Pae, 2011; Pae, Greenberg, & Morris, 2012; Ravand & Firoozi, 2016).

The Rasch model focuses on three core aspects when assessing the quality of the data: the model conditions and measurement functions of the data fitting the model, having items and persons compared on a mutual scale with standard errors, and conditions regarding items and persons fitting the model (Smith & Smith, 2004). Firstly, the conditions that the model should meet and the measurement functions can be explained in terms of persons and items, as they both need to meet these requirements. Regarding the persons: Between two persons with different ability levels, the individual with the higher ability level will have a higher chance of getting items correct than the individual with a lower ability level. In terms of the items: Between two items, the easier item will have a greater chance of endorsement by persons than the difficult item, which will have a lower chance of endorsement. Secondly, the mutual scale is referred to as the person-item map, and lastly, the fit statistics are used to assess whether the items and persons fit the model (Smith & Smith, 2004).

The reasoning behind using the Rasch model as the method by which to explore the items of the ECT lies in its advantages. One of these advantages is that intrinsic to the Rasch model is that the item difficulty is separate from the person ability. This differs from Classical Test Theory (CTT) which links these aspects together and be dependent on one another. If most candidates get items correct in a test, and one uses classical test analysis to analyse the items, it would be assumed that either the test is too easy or the candidates completing the test are of high ability. This assumption is not made in the Rasch model, as these instances are clear. The person and the item are separate entities, and one can compare them. Thus, the Rasch model has item and sample independent information (Yu, 2013). This aspect alone provides great insight into the quality of the test, as one can locate the discrepancies of misfits in the data more precisely.

Another advantage of using the Rasch model is that it allows items to be compared to each other even though they are from different subsets. This allows for a smaller number of items to be used in tests (Hula, Austemann-Hula, & Doyle, 2009). Moreover, the length of measures is usually a great concern when using CTT, because it requires longer scales to produce higher reliabilities. Furthermore, CTT requires a large item bank and analyses of items are heavily dependent on large scales. The Rasch model however, does not require long scales, and it is possible for a short scale to be more reliable than a longer scale. Thus, the Rasch model allows for a smaller number of items to be used in tests (Hula, Austemann-

Hula, & Doyle, 2009; Streiner, 2010). This is, therefore, an important justification for using the Rasch model, as the ECT is not a lengthy test and this would provide useful insights in terms of the item quality instead of merely lengthening the test for better results. Furthermore, the Rasch model have assisted in examining different inductive reasoning tests, because it allows one to summarise the items in assessment measures more concisely (Keeves, 1992). Since this is a newly developed instrument, retaining well-performing items is indispensable (Harvey & Hammer, 1999). A major advantage of the Rasch model is that it is not sample dependent, which allows the interpretation of an individual's ability level and the item difficulty to be calculated independently of the data (Erguven, 2014; Van der Elst, Reed, & Jolles, 2013). Hence, the Rasch model can separate items and persons and examine them independently. CTT, on the other hand, considers the whole test and sample when examining the performance of items and persons simultaneously. CTT can be referred to as "weak models", while the Rasch model is referred to as "strong models" (Erguven, 2014, p. 27). Additionally, the process of analysing data is relatively simple when using CTT, while the Rasch model may appear relatively complicated (Erguven, 2014).

Other considerations regarding the Rasch model are that larger sample sizes (10 persons per data point) are required, complicated mathematical analyses are conducted on the data, several options for assessing model parameters are available, and there are strict guidelines regarding the fitness of data and the model. These considerations are vital when conducting Rasch analysis and require one to be acquainted with the specific requirements when interpreting the output (Erguven, 2014).

When reviewing the theoretical underpinnings of the Rasch model as well as its advantages, it is apparent why it was chosen to achieve one of the study's objectives (statistically exploring the unidimensionality of items). The above theory is, however, not sufficient, as the interpretation of the output requires that one comprehend various aspects, which will be explored. For this reason, the various aspects of the Rasch model will be discussed and the guidelines used for the interpretation of Chapter 7 will also be identified.

6.7.2.1.2 Fit statistics

Fit statistics are used to assess how well the data fit the model (Van der Walt & Steyn, 2008). Fit statistics are comprised of the expected and observed responses that allow the data to fit the model. Fit statistics in the Rasch model is used to assist in the regulation of quality and the detection of the items and persons that are misfits (in other words, items that do not fit the model) and fail to add value to defining the construct. These misfit items and persons require further investigation to determine why they are not contributing to the model. The items and persons that fit the model can be presumed to be accurate in terms of the person's ability and item difficulty parameters (Smith, 2001; Smith & Smith, 2004).

If items fall within the appropriate range then they are considered a good fit for the model, however, when they fall outside of this range, they are identified as 'misfitting' or 'over-fitting' (Van der Walt & Steyn, 2008). According to Bond and Fox (2007), the range of 0.70 – 1.3 can be considered as the appropriate range for items that fit the model (Baghaei & Amrahi, 2011; Linacre, 2011; Pensavalle & Solinas, 2013). Items that have scores lower than 0.70 would be considered 'over-fitting', which implies that there is too little variance identified in the item. Items that have scores higher than 1.3 are considered 'misfitting', which means that the item is less able to predict performance (Bond & Fox, 2007; Pensavalle & Solinas, 2013; Van der Walt & Steyn, 2008). Furthermore, items that have infit values less than 1 indicate that there is an absence of randomness or noise observed, while items that have infit values larger than 1 suggest that there is abundant variability in the data (Smith et al., 2003).

The concept of misfit is described as problems within the data that cause either the items or the persons not to fit the model. Four misfit indices allow one to assess whether the items or persons fit the data. These four indices are infit standardised residuals (IN ZSTD), outfit standardised residuals (OUT ZSTD), infit mean squared (IN MNSQ), and outfit mean squared (OUT MNSQ) (Smith et al., 2003; Yu, 2013). When items do not fit the model, they need to be omitted or revised because these items are possibly measuring other constructs, therefore construct-irrelevant information is present (Baghaei & Amrahi, 2011; Ravand & Firoozi, 2016). Within the misfit indices, the outfit indices are sensitive to careless mistakes (Baghaei & Amrahi, 2011), while the infit indices are sensitive to unexpected response patterns (Baghaei & Amrahi, 2011). The Winsteps program (Linacre, 2015) provides information on both persons and items fitness for Rasch analyses. The infit statistics are less susceptible to unpredicted responses of items that are not related to the latent variable. The outfit statistics are, however, susceptible to items that differ vastly from the latent variable (Smith et al., 2003).

Item fit can be observed by exploring the mean squared (MNSQ) values. The cut-off value for problematic MNSQ is 1.3, recommended by Bond and Fox (2007; Linacre, 2011). This would imply that the item does not match the other items and differs significantly from the model. This item is then considered a misfit, as it is behaving differently compared to other items in the test. This also indicates that the item needs to be investigated further, to understand why it does not fit the model. Although low MNSQ values can be considered problematic, the high MNSQ values require more concern as they threaten validity. The person fit is evaluated by examining the same four misfit indices. The misfit indices found in person thetas are associated with several explanations, such as persons who are not behaving normally (in comparison to the other persons in the data), persons who have achieved either higher or lower scores than expected (in comparison to their response pattern), or the person's pattern could indicate cheating or guessing (Ravand & Firoozi, 2016; Yu, 2013).

The standardised residual (Z-score) (ZSTD) value informs one whether the data fit the model in terms of the allocation of the standardised residuals. One should, however, refrain from removing items based on high residuals as it may only improve the distribution. The ZSTD thus provides information about the model fit and not specifically the item fitness (Linacre, 2011; Ravand & Firoozi, 2016; Yu, 2013). The appropriate range for the ZSTD is between -2 and 2 for model fit (Baghaei & Amrahi, 2011; Linacre, 2011; Ravand & Firoozi, 2016). This means that values that are not within this range are considered to be statistically different.

All the fit indices used for interpretation are crucial for understanding the item and person characteristics for the test. More so, these fit statistics provide important insight into the construct validity of the instrument, specifically within the Rasch model (Smith et al., 2003). Thus, the information obtained from the items, persons, and model fitness is imperative to test development.

6.7.2.1.3 Reliability and separation values

In this study, the reliability values were interpreted in terms of CTT, specifically because the reliability index used in the Rasch model is that of Kuder-Richardson. The Kuder-Richardson *p* value was therefore interpreted as follows: A value of .60 to .69 is acceptable for research purposes, a value ranging from .70 to .79 is acceptable for a newly developed measure, a value ranging from .80 to .89 is acceptable for an aptitude test, and a value of .90 and above is acceptable for selection purposes (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009). The person reliability index indicates the degree to which the equivalent individuals would display the same ordering of proficiency on an equivalent set of items. Item reliability index indicates that extent to which items would display the identical ordering of item difficulties on a different but similar group of individuals (Bond & Fox, 2007).

The person and item separation indices indicate how well or poorly the items and persons are spread (Linacre, 2009; Struik, 2011). A minimum separation value of 2 is considered acceptable (Baghaei & Amrahi, 2011). In addition to this, a separation value of 2 indicates that the measures are statistically significant (Linacre, 2009). When the separation value is over 2, one can trust the representivity of test items (Baghaei & Amrahi, 2011). This means that the higher the separation index value, the greater the spread of persons, while low separation values are indicative of redundant items and limited variability of persons. The item separation index also provides information about whether the items can distinguish between different levels of persons. The person separation index indicates if there are persons of various abilities or whether the persons have similar abilities. The higher separation values are indicative of a good spread of persons, while low separation values tend to indicate that persons have similar abilities and that there is not a good spread of abilities across people (Linacre, 2009; Struik, 2011).

The person separation value can also be considered a classification of the sample regarding their reliability. A low person separation value (a value less than 2) would suggest that the test might not be too receptive in separating high and low performers on the test. This may suggest that more items are required. Additionally, the item separation value informs one of whether the items are ordered in the appropriate hierarchy. A low item separation value (a value less than 2) would indicate that the person sample is too small to verify the item difficulty within the hierarchy of the test (Linacre, 2015; Ravand & Firoozi, 2016).

All these elements were considered when analysing the item and person separation and reliability values.

6.7.2.1.4 Person-item map

The person-item map shows the distribution of persons and items with regard to the mean and standard deviations (Ravand & Firoozi, 2016; Smith et al., 2003). The person-item map allows both item and person information to be measured on the logits scale. The concept of logits can be understood as the logarithm of odds (Linacre, 2012a, 2012b; Maree, 2004b, 2004c; Smith, 2001; Smith et al., 2003; Yu, 2013). This scale allows one to observe the difficulty of the items while concurrently observing the individual's performance. This scale allows one to identify the higher ability individuals and the items they answered correctly while simultaneously showing the lower ability individuals and the easier items in the test. Additionally, the Rasch model is based on the probability that an individual with a higher ability than another individual should have a higher probability of correctly answering items with greater difficulty and as such would have correctly answered items of a lower difficulty (Ravand & Firoozi, 2016; Long, Bansilal & Debba, 2014; Wright, 1997; Yu, 2013).

The logits scale allows one to measure and assess items and persons comparably. Since they can be compared on a shared scale, the items and persons have equal intervals, and the issue of different attributes being measured is resolved (Smith, 2001; Yu, 2013). On the item-person map, the item difficulty is calibrated in relation to the item mean of zero and the person ability is calculated in relation to the item difficulty. The logits scale also indicates the individual's ability alongside the item difficulty, which typically ranges from 3 to -3. Lower scores (-3) implies that individuals have lower levels of ability and the items are easy, while higher end scores (3) implies their ability levels are higher and the items are difficult (Dunne, Long, Craig & Venter, 2012; Van der Walt & Steyn, 2008). The relationship of the item difficulty and person ability on the person-item map is such that where a person is aligned with an item; the person has a 50% probability of answering that item correctly. This means that if an individual's ability location is below an item difficulty level, then the individual has less than 50% probability of correctly answering the item, but if the item difficulty is below the individual's ability location, then the individual has a greater than 50% probability of correctly answering the item (Dunne, Long, Craig & Venter, 2012; Long, 2011).

The person-item map allows one to view the descriptive nature of items, in that it serves to illustrate the content facet of the construct validity of the items. This is apparent in terms of their positioning on the map. The items that have no candidates alongside them indicate problems as the ability level is not being tested by the items, which could either be below or above the ability of the persons. The gaps in the distribution of the person-item map indicated that items are required to suitably test the ability of the persons (Baghaei & Amrahi, 2011; Ravand & Firoozi, 2016).

These guidelines on interpreting the person-item map were followed when the Rasch analysis was run for both test versions of the ECT.

6.7.2.1.5 Dimensionality

Rasch factor analysis is also referred to as dimensionality, which involves analysing the residuals to find a common variance in the data that is not explained by the primary Rasch measure. Dimensionality needs to be explored in a three-step process. The first step is to check for negative bi-serial correlations and identify any problematic items. The second step requires the fit statistics, which are needed to assess the item and person misfits. The third step involves Rasch factor analysis to explore the dimensionality of the test (Bond & Fox, 2007). In addition to this, unidimensionality can be indicated by observing the first dimension having a larger than 20% of variance explained. The dimensionality of the Rasch residuals is explored by use of principal components analysis (Pae, 2011; Pae, Greenberg, & Morris, 2012).

If a definite additional dimension is found, then this dimension needs to be separated from the other dimension. When two prominent dimensions in a test become evident, these dimensions need to be separated and form two different measures (Bond & Fox, 2007). The decision made regarding an additional dimension being observed is based on considering the size and nature of the dimension. This decision is also based on several considerations regarding the applicability of the additional dimension. The output of the factor plot is comprised of standardised residuals. This plot allows one to observe the item measures of the principal components analysis and the factor loadings of the residuals. This plot assists in evaluating whether these items are difficult or are considered easy. The items that have values larger than zero must be investigated (Bond & Fox, 2007).

The process identified in the literature is the method that was used to explore the dimensionality of the ECT for both versions. The three-step process was applied and the ECT was assessed to determine if there was a strong enough additional dimension present.

6.7.2.1.6 Item characteristic curves

An item characteristic curve (ICC) shows the relation between the probability of endorsing an item and the location of the person showing the amount of trait. The shape of the curve (ICC) is important and is called a logistic function (Streiner, 2010; Yu, 2013). The ICC allows one to place candidates’ responses to dichotomous variables into a curve. The calculation of the ICC assumes that all the candidates completing the test have various levels of the ability. The ICC therefore indicates the probability of an individual of a particular ability correctly answering an item of a certain difficulty. This curve then displays the possibility of an individual with a certain level of ability choosing a particular response, which also serves to indicate their ability level. When examining the curve, the higher levels of ability will be towards to the right-hand side of the *x*-axis, while lower levels of ability will be towards the left-hand side of the *x*-axis. In addition to this, the ICC indicates model fit when the theoretical curve and the observed proportions are parallel (Erguven, 2014; Long, Bansilal & Debba, 2014; Long, 2011; Thorpe & Favia, 2012).

The ICC offers one the opportunity to observe the item difficulty level as well as whether the item can discriminate. Item difficulty is usually indicated as *b* and is situated within the ability levels of the scale. The difficult items are situated among high ability individuals while the easy items are situated among the low ability individuals. The location of an item, therefore, determines its difficulty level; to the left on an ICC curve would imply easy, and to the right would imply difficult items (Smith & Smith, 2004; Thorpe & Favia, 2012).

When an item does not discriminate sufficiently in the ICC, this will be indicated by the observed proportions which will be flatter (or smoother) than the theoretical curve. This is referred to as under-discrimination or under-fit. This means that individuals with lower proficiency performed better on an item than expected, which implies that individuals with higher proficiency are incorrectly calculated to performing on the item as if it was easier than it is in reality. This is largely due to the relationship between the item difficulty and person ability. Another item issue referred to as over-discrimination or over-fit in the ICC, is when the observed proportions are steeper or sharper than the theoretical curve. The discrimination of the item is therefore higher than expected which means that individuals with low proficiency are disadvantaged while individuals with higher proficiency benefit (Long, Bansilal & Debba, 2014).

The difficulty index (*b*) is calculated by assessing the point of the curve that crosses at the 0.5 probability value on the *y*-axis. Item discrimination, on the other hand, is depicted by *a*, and indicates if the item can distinguish between candidates of low and high ability. This is observed by evaluating how steep the curve appears (Thorpe & Favia, 2012).

The way in which the ICC is interpreted is the same for the test characteristic curve, in that it is a singular curve that represents all the item characteristic curves of the test (Yu, 2013). The Rasch model does not allow the ICC to cross, but to run horizontally to each other. This shows that unidimensionality has not been violated (Smith & Smith, 2004). In this study, the ICC was run for both versions and interpreted according to these guidelines.

6.7.2.1.7 Item and test information function

Rasch analysis uses the mathematical formula, R.A. Fischers ($I = 1/\sigma^2$), to calculate the amount of information the ICC can provide. This information is derived from how accurately the parameter was estimated. This information function can be calculated for individual items, known as item information function (IIF), and for the test, known as the test information function (TIF). The TIF is comprised of all the IIFs of the test. The TIF informs one of the accuracy of all the test parameters combined (Yu, 2013).

Additionally, the IIF can be used to establish the amount of information indicated by items along the levels of theta. This refers to the item discrimination and informs one of the measurement precision. The sum of all these IIFs comprise the TIF, which indicates how accurately the test is measuring at all levels of the construct (McBride, 2001). It should also be noted that Rasch analyses use IIF and TIF, similarly to how the standard error of measurement is used in CTT. The test information notifies one of how accurately the ability levels are estimated by the test. The test information is formulated independently from the various ability levels (Thorpe & Favia, 2012). This information is instrumental in understanding how the test is functioning and crucial in the development of the test. These guidelines were used when the TIF was run and interpreted.

6.7.2.1.8 The procedure of Rasch analyses

Three important considerations were noted when the Rasch model output was evaluated. Firstly, the model fit statistics indicated how well the items fitted the unidimensional Rasch model. Secondly, the assumption of unidimensionality was assessed by evaluating the principal component analysis of residuals and item-fit statistics. This informed the process by assessing if a secondary dimension was present. Thirdly, this study used raw data, which necessitated the use of calibrated data (logits) (Green & Frantom, 2002).

Winsteps was used to conduct Rasch analyses (Linacre, 2009). Since the ECT has dichotomous items, the Rasch model, was used. The procedure that was used to conduct the Rasch model was as follows: A control file was created in Winsteps (Linacre, 2009 Linacre, 2012a; Maree, 2004a), which specified the model parameter (the item difficulty parameters), data structure, and output format. This was done for both test versions of the ECT. Once this syntax (control file) was created, it was run using the Winsteps software (Linacre, 2009), which provided the different sets of outputs for interpretation.

The output from the syntax was interpreted as follows (Bond & Fox, 2007; Green & Frantom, 2002; Linacre, 2009 Linacre, 2011, 2012a; Maree, 2004a): The model infit and outfit statistics needed to be 1.30 for acceptable model fit, the mean infit and outfit ZSTD needed to be 0.00 for acceptable fit, and the individual item-fit statistics and person-item total correlations were assessed for unidimensionality according to the guidelines stated earlier (section 6.7.2.1.2) in the chapter.

The analyses done on the ECT version 1.2 needed to consider the effect of the time limit imposed, which made it a performance test, as individuals were expected to answer questions within a given amount of time. This relates to “test-wiseness” (Nell, 2000, p. 64), which refers to an individual’s readiness for a test environment and his or her ability to be motivated and to respond both accurately and within the given time period (Nell, 2000). Bearing this in mind, the analyses that were to be conducted on the ECT version 1.2 would have used the appropriate Rasch analysis that is used for tests with time limits to statistically control for this confounding variable (test-wiseness or test anxiety) (Verhelst & Jansen, 1992). This decision was, however, reconsidered as the same the Rasch analysis technique was used for both test versions to compare the results.

6.7.2.1.9 The reporting of Rasch analyses

To assist in the interpretation of the Rasch analysis output, the results were reported as follows:

1. The Person Statistics
2. The Item Statistics
3. Test Empirical Randomness
4. The Summary of Category Structure Statistics
5. The Person-Item Map
6. The Measure Order Statistics
7. The Bubble Chart
8. The Misfit Order Statistics
9. The Variance Decomposition of Observations
10. The Standardised Residual Contrast Plots
11. The Standardised Residual Loadings
12. The Test Characteristic Curve
13. The Item Characteristic Curves

14. The Test Information Function Curve

These results were presented separately for the two test versions.

6.7.2.2 Confirmatory factor analysis

CFA can be used to confirm factor structures within the data as well as build on the construct validity of the test. Furthermore, it is a robust and undeviating method of assessing constructs or dimensions within a test (Pae, 2012). When conducting CFA, there are a few considerations that need to be considered. These include the following: the theory that needs to be examined, the appropriate size of the sample, the measurement device, multivariate normality, specifying the parameters, missing values, analysis of model fit, and anomalies (Suhr, 2006). These are important considerations and were identified within the study.

CFA (of the ECT version 1.3) was performed based on prior knowledge of the underlying structure, which was observed in the exploratory factor analyses (EFA) previously conducted on the ECT version 1.2 (Arendse & Maree, 2017). It is a requirement that CFA be based on a strong theoretical framework or empirical knowledge. The dimensionality of the ECT will, therefore, be confirmed by conducting CFA as a form of structural equation modelling (SEM) in IBM AMOS (Arbuckle, 2010). The strongest structural confirmation would result from establishing that the same model exists between the ECT version 1.2 and 1.3. This will allow one to confirm that the structures of both versions are inherently similar.

SEM has the following advantages: It allows CFA to be calculated while lessening the error in the measurement, it allows the model to be evaluated instead of the coefficients to be evaluated independently, and non-normal and missing data are handled easily (Garson, 2015). These advantages are the reason for conducting CFA within the SEM framework.

IBM AMOS software (Arbuckle, 2010) allows for a graphical interface in which several linear models are fixed into a unifying framework that additionally reduces measurement error (Division of Statistics & Scientific Computation, 2012; Milfont & Fischer, 2010). For this reason, IBM AMOS (Arbuckle, 2010) was the preferred software to conduct CFA. This graphical input assists in visually displaying the stipulated model, which is essential when evaluating the dimensions of the constructs being assessed.

6.7.2.2.1 The theory of confirmatory factor analyses

When SEM is used to conduct CFA, it involves the use of goodness of fit indices to ensure that the model stipulated is congruent with the structures, variances, and covariances. It should, however, be noted that there may be other models that fit the data and may fit better than the stipulated model. This can, however, only be examined when explored in the analysis. This emphasizes the importance of the model that the researcher wishes to evaluate. It requires theory and discernment, as well as identifying the direction of the stipulated relationship that may affect whether a model will be accepted and be a good fit (Garson, 2015; Milfont & Fischer, 2010). For this reason, there are several aspects within CFA that need to be discussed, such as the CFA model, the variables used in SEM, the process of conducting a CFA using SEM, the graphical diagrams used to create the model, and the model fit statistics employed in CFA.

6.7.2.2.2 The confirmatory factor analysis model

A standard CFA model usually consists of two factors with six indicators. The indicators are continuous variables that claim to measure the factors and other elements, which are symbolised by the error term. The arrows used to indicate an effect on the indicators of the factor are referred to as pattern coefficients (or factor loadings) and are analysed as regression coefficients. When indicators are hypothesised to be the cause of underlying constructs, they are called effect indicators or reflective indicators (Kline, 2011). Moreover, in standard CFA models, the factors are regarded as exogenous, while the indicators are endogenous. This symbolises reflective measurement. The unique variance in the analysis is indicated by measurement error, which explains the variance that is not accounted for by the factor. The unique variance is either due to random errors or systematic variance that is not attributed to the measured factors (Kline, 2011).

In CFA, unidimensionality is expressed as a single factor that has all the indicator loadings on it and the error terms would be separate. The single factor loadings referred to as a restrictive factor model. Multidimensionality, on the other hand, is observed when any of the indicators load on to more than, or are equal to, two factors and when the error terms of an indicator co-varies with another indicator. The unidimensional models allow for better assessment of convergent and discriminant validity compared to other models. Furthermore, the standard CFA models are identified when either the single factor has a minimum of three indicators (unidimensionality is established) or two or more factors have a minimum of two indicators for each factor (multidimensionality is established). This is referred to as the three-indicator rule and the two-indicator rule respectively (Kline, 2011). Since the ECT is theorised to be unidimensional in nature, the three-indicator rule would need to be applied for the model to be identified and considered acceptable.

6.7.2.2.3 The variables used in SEM

SEM requires a strong theoretical foundation so that the model can clearly identify the latent construct and the indicators that need to be evaluated. A minimum of one or two indicators should be used, which would require considerable certainty regarding the theory. Alternatively, the use of three or more indicators is accepted. It is recommended that the indicators have pattern coefficients on their latent factors of 0.7. This may, however, be very rigid, and path weights from latent variables to indicator variables should have 0.7 as a criterion (Garson, 2015).

The latent variables consist of either or both exogenous and endogenous variables. These latent variables are unobserved factors that are measured by the indicators. Exogenous variables usually have covariance arrows indicating their relationship to other exogenous variables. They are independent variables and have previous causes or relations. Endogenous variables, on the other hand, are dependent variables and have causal relationships that use regression paths. The exogenous or endogenous variables can be responsible for the path leading to the endogenous variables (Garson, 2015). It should also be noted that manifest variables are the variables observed in the data, while latent variables are the unobserved variables in the data (Kline, 2011).

6.7.2.2.4 The SEM process

According to Kline (2011, p. 92), the following steps are taken when performing SEM:

1. Specifying the model.
2. Model identification.
3. Decide on the measures in terms of the constructs and data preparation.
4. Model estimation, which includes assessing model fit.
5. Interpreting the parameters estimation and consider similar models.
6. Re-specify the model.
7. Report on the results.

6.7.2.2.4.1 Model specification

Model specification is the initial process of performing SEM, and involves clarifying the model and exploring its associated variables. This process also involves deciding on a regression path, covariance arrows, and or the use of a particular value for either the regression path or covariance (Garson, 2016). The model specification represents the theorised model in a graphical input, which allows one to define the parameters regarding the manifest and latent variables through the use of sample data. Underlying the structural model in the graphical input are the various equations that allow one to establish whether the theorised model is acceptable (Kline, 2011). There are, however, a few aspects that need to be considered for CFA to be considered acceptable when specifying the model. These are: that the model is theory-driven, the number of factors required, the items that belong to the factors, and the error associated with the model (Suhr, 2006; Suhr & Shay, 2009).

This step, model specification, is crucial to the model being eventually accepted, as the steps that follow assume that the specified model is accurate. It should also be noted that when specifying the model, it can be based on empirical evidence or a theoretical framework (Kline, 2011). The process of model specification includes using theory or theoretical information, specifying the model using diagrams, identifying the model, making parameter estimation for the model, examining model fit, and reporting on results (Suhr, 2006). The process of model specification is therefore repeated if the model is not accepted.

6.7.2.2.4.2 Model identification

Model identification involves the model having particular estimates for the parameters identified. The identification pertains to the structure of the model and is not dependent on the sample data. If the model is not identified, then the model will not be analysed by the SEM program. The identification of the model is dependent on the SEM that is used, and must adhere to certain criteria for the model to be deemed satisfactory. The structural model consequently needs to comply with two requirements: The degrees of freedom for the model must minimally be zero, and a scale must be attached to all latent variables being measured.

This criterion assists the model in being identified by the SEM program. If the model is, however, not accepted, then it needs to either be adjusted, or a new, better fitting model needs to be generated. Thus there are alternatives available for model testing when a model is not accepted (Kline, 2011).

SEM typically requires a large data set, as the sample size is linked to the complexity of the model being evaluated. A recommended ratio of sample size to parameters is 20:1. Most studies have reportedly used 200 cases for SEM, yet this may be problematic if the model is highly complex (Kline, 2011). When the data have already been collected (secondary data) for the study, the sample size may pose a particular problem with identification of the model. This can, however, be solved by increasing exogenous variables and indicators, which would assist in identifying the structure of the model and the measurement of the model (Kline, 2011). This was, however, not a problem associated with the current study, as both sets of data were well over the recommended sample size of 200.

6.7.2.2.4.3 Model estimation

Model estimation is regarded as the analysis of the model and requires several aspects to be examined. Firstly, the model fit needs to be assessed. Model fit implies that the data correctly fit the model that was specified. It is often found when examining model fit that there is a need to re-specify the model, as the data do not fit the model. When re-specifying the model, there should still be reliance on theory or empirical knowledge. Secondly, once the model fit is acceptable, the parameter estimates need to be evaluated. This requires one to establish whether the parameter estimates are providing valuable information. Thirdly, the consideration of other relevant models that may fit the data should be explored. The exploration of possible models that may fit the data is essential in arguing why a specific model should be chosen above other models. This also assists in the validity argument for the model chosen (Kline, 2011).

A commonly used statistical method of estimation in SEM is maximum likelihood. This statistical estimation method has many advantages over other estimation methods (like the two-stage least squares) as it can handle complex data sets, operate effectively, and be reliable (Kline, 2011).

6.7.2.2.4.4 Model fit statistics

There are a few statistics that need to be reviewed to establish if the model fit is satisfactory (Milfont & Fischer, 2010; Suhr, 2006). Moreover, when evaluating the goodness of fit indices, the complexity of the model needs to be considered (Suhr, 2006). The goodness of fit statistics assesses how much of the covariance in the data the model explains. Essentially, it assesses whether the specified model fits the data better than not having a model (Kline, 2011). The first statistic to be considered is the chi-square statistic, which is a goodness of fit statistic (Milfont & Fischer, 2010). The chi-square model statistic assists in evaluating the fit of the model by considering the covariance matrix and the data fitting this covariance matrix in light of the model identified (Kline, 2011). The chi-square statistic can be used for hierarchical models, while Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) can be used for non-hierarchical models (Garson, 2016). Essentially, the chi-square statistic allows one to establish the relation between the expected and observed covariance matrix. The chi-square value can range from 0 to 1, but needs to be close to zero to indicate less difference, which is what one intends to find (Suhr, 2006).

A non-statistically-significant chi-square test statistic suggests that the model fits the covariance data. One is, however, cautioned that this merely reports on the covariance aspect and not whether the model is specifically accurate. This implies that the model needs more inspection than simply accepting the significance (*p* value) of the chi-square statistic. There are also several limitations regarding the chi-square statistic, such as its sensitivity towards severe non-normal distributions, large correlation sizes, high quantities of unique variance, and large sample sizes, which can contribute to incorrect interpretations being made (Kline, 2011). Since the chi-square statistic is susceptible to sample size, it cannot be the only statistic used for assessing model fit as it will not be sufficient (Milfont & Fischer, 2010).

For this reason, other statistics are utilised to explore model fit and are grouped within either absolute or incremental fit statistics. Absolute fit statistics aim at assessing the theorised model fit alongside the sample data. Within the absolute fit indices, the chi-square statistic, the Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR) are considered. These absolute statistics are considered acceptable based on specific criteria. Incremental fit statistics compare the model and evaluate improvements in the model fit. Within the incremental fit indices, comparative fit index (CFI), and Bozdogan's consistent version of Akaike's information criterion (CAIC) are considered. The incremental statistics require lower values for the model to be considered a good fit (Milfont & Fischer, 2010).

The second statistic that was explored for model fit is the RMSEA. The RMSEA statistic is an important model fit statistic for CFA models and examines the residual in the model (Suhr, 2006). Moreover, the RMSEA statistic assists in assessing whether the specified model fits the population. It may appear biased in models with small degrees of freedom (a higher RMSEA value will be observed) (Morgan, 2015). Additionally, the RMSEA indicates unexplained variance, and it should, therefore, be as small as possible (Hoyle, 1995; Hu & Bentler, 1999). The RMSEA value ranges from 0 to 1 (Suhr, 2006). This statistic thus assists in assessing the model fit, and the closer the value is to zero, the better the model fit (Suhr, 2006; Kline, 2011).

The third statistic that was used to explore model fit was the CFI. The CFI statistic assesses the model fit by evaluating the specified model alongside a standard model and assesses whether the specified model would improve the model fit, in comparison to a standard model (Kline, 2011). The CFI statistic also explores the sample size adequacy. The CFI value ranges from 0 to 1, where a value closer to 1 implies best fit (Suhr, 2006).

Once the model fit indices are shown to indicate satisfactory model fit, the parameter estimates need to be explored. The standardised parameter estimates are examined instead of the unstandardized parameter estimates (Suhr, 2006), as they are considered more accurate.

6.7.2.2.4.5 Model diagrams

The diagrams and symbols that were used in the analysis and graphical input of the model are as follows (Kline, 2011; Milfont & Fischer, 2010):

⬭ : this indicates the unobserved latent variable in the model.

▭ : this indicates the observed manifest variables in the model.

○ : this indicates the indicators of the model.

→ : this represents the direct effect of one variable on the other variable.

↔ : this indicates covariances as correlations between variables.

6.7.2.2.5 The procedure of confirmatory factor analyses

The AMOS graphical input was used to graphically plot the latent variable, manifest variables, and indicators. This software made it easy to construct a model. For the ECT, the empirical model established in the EFA analysis (Arendse & Maree, 2017) was tested to confirm if this model best fits the data. Present in both models was the latent variable of verbal reasoning, as it was hypothesised that this variable is the underlying variable present in the factors of the ECT.

The execution of CFA involved defining the model, which was previously identified by the EFA; displaying the model in the graphical format; selecting the outputs; and calculating the estimates. Thereafter, the model fit statistics were observed and evaluated to establish whether the model could be accepted. Once the model was accepted, the parameter estimates were explored.

The criteria used to evaluate the chi-square test for this study was that the value observed should be close to 0 and the probability value should be smaller than 0.05 (Hoyle, 1995; Hu & Bentler, 1999). Thus, a chi-square statistic that was not statistically significant was the criteria by which the model was accepted. The criterion used to evaluate the RMSEA statistic was that the value needed to be smaller than 0.06 (Hoyle, 1995; Hu & Bentler, 1999). The criteria used to evaluate the CFI, non-normed index, normed fit index, and Tucker-Lewis index was that the values observed needed to all be greater than 0.9 (Hoyle, 1995; Hu & Bentler, 1999).

The squared multiple correlations, the standardised regression weights and the baseline comparison were also interpreted to provide information on the model. These statistics were essential in determining the existence of an acceptable model for the ECT.

6.7.2.2.6 The reporting of confirmatory factor analyses

To assist in the interpretation of the CFA output, the results were reported as follows:

1. The Graphical Input of the ECT Model
2. The Assessment of Normality
3. The Regression Weights
4. The Squared Multiple Correlation
5. The Standardised Regression Weights
6. The Model Fit Statistic (Chi-Square Statistic)
7. The Baseline Comparison
8. The Root Square Error of Approximation (RSMEA)
9. The Root Mean Square Residual (RMR)
10. The Akaike's Information Criterion (AIC)

These results were presented for only the one test version (ECT version 1.3).

6.7.2.3 Multi-trait multi-method analysis

Multi-trait multi-method (MTMM) is a method used to explore construct-related information. MTMM is based on the notion that the construct does not need to be linked to any specific method, or be linked to any other irrelevant construct. This allows one to explore the construct relationship. The relationship between the same constructs should be high and should not be influenced by the method used. MTMM requires one to explore coefficients for reliability, convergent, and discriminant validity. Convergent validity requires a correlation between the same or similar constructs which are obtained from two different methods of assessment (Goodman, 2004; Palmer & Bachman, 1981; Yun & Ulrich, 2002). Discriminant validity requires correlations with different constructs that were obtained by using two different methods of assessment (Goodman, 2004; Palmer & Bachman, 1981; Yun & Ulrich, 2002). MTMM is interpreted by exploring the reliability and validity coefficients. The reliability coefficient needs to be larger than the convergent and discriminant validity coefficients. The discriminant validity coefficient should be smaller than the convergent validity coefficient, and the mono-trait hetero-method coefficients should be larger than the hetero-trait hetero-method. The reliability and validity coefficients provide construct related evidence and allow one to create a powerful argument for construct validity (Palmer & Bachman, 1981; Yun & Ulrich, 2002).

The use of MTMM analysis was required as a means of presenting evidence of the construct. This technique assists in providing proof of construct validity in the form of convergent and discriminate validity. Validity is established by the use of correlations to form the matrix. In the USA, the use of MTMM is encouraged as it serves to fulfil part of the requirements of the Standards for Educational and Psychological Testing and the No Child Left Behind Act (Reyes & Johnson, 2010). Furthermore, MTMM is useful when assessing performance on language tests, specifically when diverse populations are tested. The exploration of methods and traits allows for a comprehensive look at the construct and how it unfolds in the assessment (Pae, 2012).

6.7.2.3.1 The theory of the multi-trait multi-method analyses

Campbell and Fiske (1959) created the MTMM as a means by which to examine construct validity. The process of examination is divided in terms of traits and methods. The traits refer to abilities, while methods refer to the means by which these abilities are examined, such as a testing format. There are several arrangements of traits and methods that allow one to examine the convergent and discriminant validity. The matrix containing the correlations of the variation in traits and methods needs to be interpreted in a specific way. Convergent validity is established when the correlations are high for hetero-method mono-traits, while correlations are low for mono-method-hetero-traits. Discriminant validity is established when there are lower correlations between different traits and hetero and mono-methods (Maas, Lensvelt-Mulders, & Hox, 2009).

MTMM allows one to explore several types of correlations regarding their relation to one another, such as those with similar constructs and those with similar methods. The use of opposing constructs and diverse methods creates an opportunity to evaluate similarity and dissimilarity. Campbell and Fiske (1959) identified four criteria, one of which refers to convergent validity, while the remaining three support the identification of discriminant validity. The first criterion relates to the mono-trait hetero-method, which is also referred to as measures assessing the same construct and using different methods to do so. There should be a strong correlation which would imply that there is a substantial relationship due to the variance explained. The second criterion indicates that the mono-trait hetero-methods correlations need to be larger than the correlations observed by hetero-trait hetero-methods (which would be found in the rows and columns of the matrix). The third criterion refers to mono-trait hetero-methods, which values need to be higher than the values obtained by the hetero-traits mono-methods. The fourth criterion refers to the same trend being observed for the correlations across traits, regardless of mono-method or hetero-methods (Goodman, 2004).

MTMM has been used extensively in language testing, due to its ability to examine both traits and methods as well as how these may impact learners’ performance. Furthermore, MTMM is based on the fact that there are more high correlations present in mono-trait than mono-method. Hetero-trait correlations should be lower than mono-trait and hetero-methods correlations should be lower than mono-methods correlations (Pae, 2012).

The advantage of using the traditional MTMM analysis is that methods and traits can be examined separately. Another advantage is that it allows reliability and construct validity coefficients (which include convergent and discriminant validity) to be indicated. Convergent validity requires correlations between the same constructs utilising various methods, while discriminant validity requires correlations between various constructs utilising the same methods (Schumacker & Lomax, 2010). These advantages are why MTMM was chosen as the preferred method for exploring the correlations of similar and different constructs with the construct (verbal reasoning) of the ECT.

6.7.2.3.2 Limitations of using the original MTMM

The MTMM design has been considered effective, yet the disadvantages associated with this design are worth noting. Firstly, there is vagueness regarding the acceptability of the results obtained from the MTMM analysis. Secondly, the uncertainty on removing method and trait factors within the correlation matrix is also considered problematic. Based on the disadvantages associated with the original MTMM design, the currently most preferred method of conducting MTMM analyses is to use SEM (Langer, Wood, Bergman, & Piacentini, 2010).

Other criticisms levelled at the technique are that the reliance on correlations for both traits and methods can cause confusion and unclear guidelines. The evidence of sufficient correlations for either mono or hetero-traits is indefinite and can lead to vague adherence to the intended aims of MTMM. MTMM is limited in its use of correlations, as one cannot unquestionably claim them to be dimensions, but can only claim that there is a relation between the constructs. Another limitation of MTMM is that it is not able to distinguish random error from method variance. Since there are no descriptive criteria that stipulate the minimum required size of the correlation for a significant difference to be indicated, this gives one range in which to either argue for or against what is perceived as low or high correlations (Pae, 2012).

Another disadvantage associated with the original MTMM model is that it does not allow for easy analysis. Based on this, Schumacker and Lomax (2010) advocated for advanced techniques, such as SEM, to be used when conducting MTMM analysis. SEM has been demonstrated to be more useful than the original MTMM analysis (Schumacker & Lomax, 2010). This suggestion would, however, apply to data that are more applicable to the analysis (when multiple traits and methods are available for analysis) and was consequently not considered for the ECT data sets.

Another criticism levelled at the original MTMM method is the subjective inferences that are made from the analysis. Due to these criticisms, specific developments were implemented to improve MTMM analyses, specifically to limit its subjective nature and provide more guidance regarding convergent and discriminant validity being observed. These advancements were in the form of the following methods: analysis of variance and CFA models, which included the correlated traits-correlated method; correlated trait-correlated uniqueness; the correlated traits-constrained uncorrelated method model; the correlated traits-constrained method; and correlated traits-uncorrelated methods (Lance & Fan, 2016). All these advances, however, rely on multiple methods, which are a limitation for the current study, as the constructs being compared have the same method.

Since SEM was used in the CFA for the factor structure of the test, it was not necessary for the purposes of the current study to conduct SEM for the MTMM analysis. A limitation regarding the use of SDA is that the data limit the use of more advanced methods. For this reason, the MTMM conducted in this study can be viewed as a correlation matrix, with a specific focus on constructs (indicated by the mono-method triangles).

6.7.2.3.3 The procedure of the multi-trait multi-method analyses

Information on the construct validity of the ECT was established by providing evidence of convergent and discriminant validity using the original MTMM analysis (Campbell & Fiske, 1959). This method of analysis seeks to confirm the existence of the same or a similar construct between two different instruments, while confirming the absence of the construct in other instruments (Campbell & Fiske, 1959; Koch, 2013). The MTMM was executed with the use of Pearson correlations (Cohen, 1988). According to Cohen (1988), the ranges by which the magnitudes of the relationships are evaluated are: 0 – .29 (small correlation), .30 – .49 (moderate correlation), and .50 – 1.00 (large correlation).

Since the tests that were used in this study are all psychometric tests and secondary data were used, there are certain limitations that prevent one from completing a generic MTMM approach. For this reason, a modified multi-trait mono-method approach was used to achieve the objective of obtaining supporting evidence of construct validity. The reason for using mono-method is that all the methods used in this study are the same (all psychometric tests) and thus hetero-methods are not applicable as they were not collected during the data collection process. The correlations that were interpreted were mono-trait (similar traits) and hetero-trait (opposite traits) as these were the constructs being measured in this study.

This modified multi-trait mono-method approach, therefore, consisted of correlations relating to the mono-trait mono-method and hetero-trait mono-methods. These were compared and evaluated to determine if there was sufficient evidence for convergent and discriminant validity. These correlations were compared in the MTMM matrix, which uses different diagonals, triangles, and blocks to sufficiently compare correlations. The diagonals, blocks, and triangles could, however, not be used in the matrix. For this reason, only the reliability diagonal was inserted in the matrix. The mono-trait mono-method triangles were interpreted separately to limit confusion. The reasoning for separating these elements from the traditional MTMM matrix was that these elements are only possible if hetero-methods are also included. In addition to this, the exclusion of hetero-methods will not allow the matrix to make sense as all the traditional elements will not be in the matrix. The modified matrix consequently only used the reliability diagonal (mono-trait mono-method), hetero-trait mono-method triangle, and mono-method blocks, because the others (such as the validity diagonals, hetero-trait hetero-method triangle, and hetero-method blocks) were not applicable.

In terms of the four criteria identified by Campbell and Fiske (1959), the analysis of the ECT data does not allow for all four criteria to be met due to the limitations relating to using SDA. Since the ECT data were collected before the analyses were run, and the same type of method was employed throughout (psychometric pen and paper tests), the MTMM for the ECT is inherently limited in its analyses. This will, however, not influence the study as the CFA will address the shortcomings associated with employing the modified MTMM. The fourth criterion is the only criterion to which the ECT data and related test data adheres. As a means of addressing these criteria, an adjustment will be made to the four criteria to superficially assess if convergent and discriminant validity is established.

The adjusted first criterion for this study was addressed by exploring the mono-trait and mono-method correlations, as no hetero-method correlations were available. The identification of large, substantial correlations was still adhered to. The second criterion was addressed by exploring the mono-trait mono-method correlation, which needs to be larger than the hetero-trait mono-method correlation. The third criterion was attended to by exploring the mono-trait mono-methods, which were explored in the adjusted second criterion as hetero-methods were not available. The fourth criterion was addressed by focusing on the trend observed across the mono-traits and hetero-traits for mono-methods. The method is, however, not the focus in this criterion.

In order to establish convergent and discriminant validity, the ECT was correlated with the following tests: AAT 1, AAT 2, AAT 3, AAT 4, AAT 5, DAT 2, DAT 3, DAT 9, DAT 10, SAT 2, SAT 4, SAT 5, SAT 6, SAT 7, SAT 8, and SAT 10. These tests were correlated with the ECT and each other and formed a matrix. The correlations between the verbal reasoning tests and the ECT should be higher than the correlations between the ECT and other constructs, as this will indicate convergent validity. The correlation between the ECT and the calculations and spatial tests should be very small, and should be much lower than the correlations observed with the verbal reasoning tests for discriminant validity to be established.

6.7.2.3.4 The reporting of the multi-trait multi-method analyses

To assist in the interpretation of the MTMM analysis output, the results were reported as follows:

1. Correlations of the ECT and the Different Psychometric Tests (AAT, DAT, and SATs)
2. The Reliability Diagonals Within the Psychometric Tests
3. The Mono-Trait Mono-Method Triangles
4. These results were presented separately for the two test versions.

6.7.2.4 Differential test functioning

Measurement invariance tests are conducted to examine problematic items within a test, by specifically focusing on the test as a whole. Measurement invariance is evaluated by conducting a differential test functioning (DTF) analysis. The need to explore (DTF) became evident when exploring how the different items of the test functioned in the Rasch analyses.

Moreover, test development and cultural fairness emphasises that items should function similarly across cultures and if not, they should be adapted to eliminate the bias present. For this reason, a DTF analysis must be conducted to evaluate the test. DTF assesses the measurement invariance and assists in establishing if the test performance is equivalent across two different groups by using all the test items (Pae, 2011; Van de Vijver & Rothmann, 2004).

According to Rasch analyses principles, the difficulty levels for items should remain relatively constant across subsets of the sample. The invariance of the test is substantiated when the difficulty level is not constant across sample subsets. Moreover, the test should not function vastly differently across the subgroups. The invariance principle is violated when items function similarly across groups. Thus invariance is attributed to dissimilarity across groups.

Invariance may imply that the test is biasing a particular group and the group's performance cannot be compared as a result (Pae et al., 2012). The importance of evaluating the test for measurement invariance stems from addressing societal inequalities. Since the ECT will be used in a multicultural context, the need to assess invariance across groups is crucial to further test development.

Based on the research question, the type of analysis that was conducted was DTF and not differential item functioning (DIF). Although DIF is necessary to perform in the construction of a test, the concern of the current study is not on specific items. Focusing on the specific items would be a useful recommendation for a follow-up study on the ECT regarding the problematic items identified in the different analyses used in this paper.

The aim of the DTF was to compare the performance of different genders and races on the ECT. Every test used and constructed in South Africa needs to be considerate of the fact that language, culture, gender, and race may influence how individuals perform on psychological tests. This increases the need to demonstrate whether these external influences impact performance on the tests as well as which items are potentially biasing certain gender or race groups. Exploring invariance is, therefore, indispensable since this sample, and the population of South Africa is multilingual and multicultural. Cross-cultural research promotes the use of DIF as a method of addressing invariance in diverse populations (Struik, 2011).

6.7.2.4.1 The theory of differential test functioning analyses

DTF occurs when a test incurs many problematic and biased items, which compromises the quality of the test (Pae & Park, 2006). DTF is also referred to as test bias and can be recognised as measurement invariance for two groups of test takers who are not performing similarly on the test (Zumbo, 2003).

The Rasch model is better suited to assess DTF than the methods used in CTT, because of the advantages associated with processing DTF using the Rasch software. Firstly, the process of assessing DTF is less complicated using the Rasch program, as it is easier to compute. This program allows the item difficulty and person ability to be explored independently. This analysis is also sample independent. This implies that the detected invariance is the difference observed only due to the person's ability and item difficulty (Bond & Fox, 2007). Since the DTF sample was split into gender and racial groups when it was analysed, the differences associated with these comparisons were attributed to the category being explored. This deviation in findings indicated which items are biased towards a particular race or gender.

The DTF that was processed in Rasch analysis requires one to consider the evaluation of the scatterplot, which plots all the items of the test. The scatterplot consisted of an identity line, which ran through the means of the two compared groups (African and White; and African and Coloured groups). One group of item difficulties or item measures is situated on the *y*-axis, while the other group of item measures is situated on the *x*-axis. The 95% confidence lines are on either side of the identity line and act as the last perimeter by which DIF items can be detected (Linacre, 2012d). Invariance should produce a constant relationship between the subsets of the sample in terms of their item difficulties. The 95% confidence line is based on the standard errors for the two groups of items. The dotted lines represent the 95% confidence lines and indicate the ideal Rasch relationship for the two groups being compared. It is expected that some would deviate from this ideal. The 95% confidence lines take the standard errors into account and allows for a fair spectrum within which the items need to fall to be invariant (Bond & Fox, 2007).

In the scatterplot, the empirical line is the line that links the observed similarity between the two sets of data, thereby creating a best fitting line through the data. The identity line, on the hand, corresponds to the normal identity line (also referred to as the slope) and passes through the origin of the two axes. In simple terms, it passes through the means of the two measures being compared (Linacre, 2015).

6.7.2.4.2 The procedure of differential test functioning analyses

The question for this study pertained to how individuals of different genders and races performed on the test, which simply put was to explore whether all the items of the ECT functioned similarly across gender (male and female) and racial (African, White, and Coloured) groups. Since there were three racial categories to be compared, the African group was the reference group as they were the largest and formed the majority of the sample (Linacre, 2012d).

The DTF analysis was conducted to establish whether there were any differences across gender groups and any differences across racial groups. Measurement invariance was explored by conducting a DTF analysis within a Rasch framework using Winsteps (Linacre, 2009). Linacre's (2012d) process of conducting a DTF analysis was followed. Firstly, the data were split into the two gender categories, males and females, in Rasch analysis using Winsteps. Secondly, the calibrated measures and standard errors were run for each category in the output table (Table 14). Thirdly, the measure values (item difficulty) for males and females were plotted and compared on a scatterplot. Fourthly, the scatterplot had an identity line drawn through the middle of the distribution and two 95% confidence lines were drawn on either side of the identity line. Fifthly, the scatterplot and a table of values relating to the item difficulty for the compared gender distribution were generated for interpretation of the DTF. The same process was applied to compute the racial categories' (African, White, and Coloured) DTF.

The DTF requires one to conduct individual analyses for males and females and then use the item difficulties of the two groups, also referred to as the item measures in Rasch, for gender comparison (Linacre, 2012d). The same was applied to the racial categories and the African group was used as the reference group between the other race categories (White and Coloured), because it formed the largest sample (group). The scatterplot was interpreted as follows: All the items that fell outside of the confidence lines were considered possibly DIF items (Linacre, 2012d).

Part of the output is a Microsoft Excel table that indicates the item's measure values, standard errors, and relevant *t* statistics. The *t* statistic is calculated as the difference between measures relative to their means (Linacre, 2012d). This *t* statistic needs to be significant at $p < 0.05$ (Pae, 2011). The empirical slope was also interpreted, and this required the value of the empirical slope to be close to the value of 1. The values for the empirical trend line and identity trend line were also observed, and it is acceptable for these two trend lines to have similar values. It would imply that the means are close to the origin. The correlation in the Microsoft Excel output was interpreted according to the ranges proposed by Cohen (1988), and indicated the strength of the relationship. These ranges are as follows: .0 – .29 (small correlation), .30 – .49 (moderate correlation), and .50 – 1.00 (large correlation) (Cohen, 1988).

The reliability values were interpreted according to CTT and the ranges used for tests, which are as follows: A value of .60 to .69 is acceptable for research purposes, a value ranging from .70 to .79 is acceptable for a newly developed measure, a value ranging from .80 to .89 is acceptable for an aptitude test, and a value of .90 and above is acceptable for selection purposes (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009). The disattenuated correlation was also observed. This disattenuated correlation is an almost perfect correlation, as it is calculated without measurement error. Thus, the disattenuated correlation will always be higher than the correlation.

6.7.2.4.3 The reporting of differential test functioning analyses

To assist in the interpretation of the differential test analyses analysis output, the results were reported as follows:

1. The Average Fit Statistics for Males and Females
2. The DTF Scatterplot with Empirical Trend Line for Gender
3. The DTF Scatterplot with Identity Trend Line for Gender
4. The DTF Statistics for the Gender comparison
5. The Average Fit Statistics for the White, Coloured, and African Race Group
6. The DTF Scatterplot with Empirical Trend Line for the African and White Race

Group

7. The DTF Scatterplot with Identity Trend Line for the African and White Race Group

8. The DTF Statistics for the African and White race groups

9. The DTF Scatterplot with Empirical Trend Line for the African and Coloured Race Group

10. The DTF Scatterplot with Identity Trend Line for the African and Coloured Race Group

11. The DTF Statistics for the African and White race groups

These results were presented separately for the two test versions.

6.7.2.5 Reliability analyses

Reliability is referred to as the absence of measurement errors in instruments (Suhr & Shay, 2009). The reliability of the ECT was calculated using Kuder-Richardson Formula 20, which can be used for dichotomous items (Suhr & Shay, 2009). This reliability statistic was used to assess how consistent the items of the test are as a whole. The value associated with the reliability for examining the internal consistency of the scale ranges from 0 – 1, and thus the closer this value is to 1, the more reliable the test will be (Mushquash & Bova, 2007; Ritter, 2010; Sabri, 2013; Weir, 2005). This value is also associated with measuring unidimensionality, as the more reliable the instrument is, the more aligned the instrument is with unidimensionality. This association is, however, not a specific trait of reliability and cannot be deduced from the reliability coefficient alone.

6.7.2.5.1 The theory of reliability

The Kuder-Richardson Formula 20 is commonly used as the indication of reliability, which either affirms or questions the internal consistency of the test. Reliability is essentially concerned with the internal consistency of instruments. For this reason, reliability coefficients are used in assessments related to psychological measures (Schmitt, 1996). A high reliability value indicates that the obtained score reflects the true score, thus there is minimal error in the observed score. This also means the measure will be consistent across different environments and will yield similar results (Erguven, 2014).

Social sciences are often burdened with the reality of unreliable measures, since scientific processes are not always followed, and as a result, the results obtained from assessments are inconsistent across different testing sessions. Such instances have created urgency in social scientists and psychologists to ensure they avoid obtaining inconsistent results and they attempt to ensure that their measures are both reliable and valid (Erguven, 2014; Tomu, 2013).

There is, however, a misperception regarding reliability. This occurs when the reliability coefficient is evaluated, and its uses are extended beyond what it can measure. Internal consistency, often measured by Kuder-Richardson Formula 20, can report on the relation between the items within the test. This does not imply homogeneity or unidimensionality, and therefore it cannot report on such instances. It is nonetheless a prerequisite to establish internal consistency when establishing homogeneity. Kuder-Richardson Formula 20 cannot, however, be used as a single measure of unidimensionality, as it can only be used to supplement other information regarding unidimensionality. Furthermore, Kuder-Richardson Formula 20 is sensitive to test length and thus the longer the test, the higher the *p* value will be (Fisher, 1992; Schmitt, 1996). Test length is a common problem with CTT, as the degree to which statistics are dependent on item and persons' numbers limits the interpretations made.

The assumption that Kuder-Richardson Formula 20 can be used to ascertain unidimensionality of items may not always be true, and thus other sources of evidence need to be acquired to validate this. A large standard error of Kuder-Richardson Formula 20 is associated with either multidimensionality or an error in sampling, which is demonstrated by the inter-item correlations. Additionally, low Kuder-Richardson Formula 20 values are attributed to short tests, because the length of a test contributes to a low Kuder-Richardson Formula 20. There are, however, instances when Kuder-Richardson Formula 20 may be low, but will still provide valuable information (Schmitt, 1996). Schmitt (1996) argues that the cut-offs used for Kuder-Richardson Formula 20 should not be inflexible, as low *p* values may still be acceptable. These considerations are important and are often forgotten when reviewing reliability coefficients. The implications associated with the value of the coefficient needs to be considered, and not merely regarded as acceptable or unacceptable.

When constructing tests, concepts such as unidimensionality and reliability are important. Unidimensionality refers to items sharing a common construct, while reliability refers to the precision of the instrument to measure this construct. Either one of these requires the other, as the instrument will lack meaning or has measurement error. Establishing both unidimensionality and reliability is thus the key to effective measurement evaluation. Unidimensionality is commonly assessed by establishing a shared factor among items, while reliability is established by assessing the value of variance and measurement error in the test (Baghaei & Amrahi, 2011; Van der Heijden, Van Buuren, Radder, & Verrips, 2003).

Outside factors may influence the degree to which one may ascertain that the measure is unidimensional. These factors affect the construct being measured and make the validation of instruments problematic. Such factors could be test-wiseness, cognitive styles, test-taking strategies, fatigue, motivation, and anxiety. These factors affect most tests, especially timed aptitude tests, and thus establishing whether the instrument is unidimensional is not clear. Although these factors are unforeseen, they still need to be considered as possible inhibitors to test validity. The Rasch model has, however, made assessing unidimensionality simpler, as the general stringent meaning of unidimensionality implies one construct (which is psychologically what one intends when assessing certain constructs), whereas the Rasch model focuses on a psychometric determination of unidimensionality, which allows one to examine this scientifically.

Unidimensionality, therefore, materialises into identifying a distinct arrangement of scores in the data (Baghaei & Amrahi, 2011). In the Rasch model, the pattern associated with unidimensionality of scores is known as the Guttman pattern. This pattern assumes that items will be distributed from easy to difficult, with high performing individuals obtaining all the easy items correct and some difficult items correct. A deviation from this pattern of answering is a deviation from the Guttman pattern and questions the unidimensionality of items (Baghaei & Amrahi, 2011).

CTT aims to establish reliability as a means of lessening the error in a test. The reliability coefficient indicates the error variance. In CTT, reliability is regarded as a property of the test data and not of the test. According to the APA regulations, the reporting of the reliability coefficient is required. The reliability coefficient may not be reported in isolation and must have accompanying information. The required accompanying information is the method in which the reliability coefficient was formulated, the sample information, and the data collection process. The reliability coefficient in CTT is, therefore, sample dependent (Erguven, 2014).

A shortcoming of Kuder-Richardson Formula 20 is that it is dependent on the population’s allocation of true scores. This would imply that Kuder-Richardson Formula 20 may be insufficient for the intended population. The use of Kuder-Richardson Formula 20 has brought about the issue of incorrect assumptions due to the *p* value (alpha) being used as confirmation of both the reliability coefficient and unidimensionality. This occurs when individuals confuse reliability, sometimes referred to as internal consistency, with assessing unidimensionality of the measure. A high *p* value can be found in cases of both unidimensionality and multidimensional items. The *p* value (alpha) is consequently not able to confirm or reflect the dimensionality of items (Van der Heijden et al., 2003). In the Rasch model, the misfit of measurement is managed by lessening the items so that they can be more unidimensional, or using two or three parameter item characteristic curves, which add parameters that improve flexibility (Van der Heijden et al., 2003). Essentially, Van der Heijden et al. (2003) argue that the *p* value (alpha) cannot be used to establish unidimensionality and that other appropriate methods be used to determine the dimensionality of items.

Since this study has used the Kuder-Richardson formula 20, the formula needs to be explained to understand how the reliability coefficient is generated. The Kuder-Richardson statistic is demonstrated as follows (Ritter, 2010; Sabri, 2013):

$$\text{KR 20} = \frac{n}{n-1}\left(\frac{SD^2 - \Sigma PQ}{SD^2}\right)$$

The components indicated in the formula above represent the following:

n = the number of items in the sample

SD^2 = the variance of the scores, also known as the square of

the standard deviation of the scores

P = the amount of correctly answered scores

Q = the amount of incorrectly answered scores

ΣPQ = the sum of the correctly and incorrectly answered scores (Ritter, 2010; Sabri, 2013).

The Kuder-Richardson Formula 20 was interpreted as follows: the range .60 to .69 is acceptable for research purposes, the range .70 to .79 is acceptable for a newly developed measure, the range .80 to .89 is acceptable for an aptitude test, and values of .90 and above are acceptable for selection purposes (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009).

6.7.2.5.2 The procedure of reliability analyses

A reliability analysis was also conducted on SPSS 23. This analysis uses Kuder-Richardson Formula 20 and provides important information regarding the scale, allowing one to improve the scale statistics. The information that was examined in the reliability analysis was the Kuder-Richardson Formula 20 *p* value on standardised items, the total item statistics, and the mean value for the remaining sample.

The reasoning behind using the total item statistics is that it informs one of the specific contributions made by the different items in terms of the scale's reliability. The items that were removed were those that negatively impacted the scale's reliability, and hence, by removing these items, the scale's reliability coefficient was improved. This does not, however, indicate that these are necessarily the problematic items within the scale. These items merely improve the scale's reliability when removed. For this reason, these items were removed, and the reliability analysis was rerun until the highest coefficient was observed. This value was then evaluated for its effectiveness. Within this evaluation, the sample size and mean were reported as they were no longer the same as the original size.

The Kuder-Richardson Formula 20 was generated in SPSS 23 using the reliability analyses option. This procedure was followed for both test versions and reported. The SPSS reliability analysis allowed one to determine the items that either improved or inhibited the coefficient value, and this is an important aspect of evaluating the reliability of the test. A revised reliability value was generated by eliminating the items that inhibited the *p* value. The elimination of these items produced an improved *p* value. This procedure was followed for both test versions and reported.

6.7.2.5.3 The reporting of reliability analyses

To assist in the interpretation of the reliability analysis output, the results were reported as follows:

1. The Reliability Results

2. The Revised Reliability Results

The results were not presented separately for the two test versions, as it was more valuable to present these values for the two test versions simultaneously.

6.8 Conclusion

The chapter focused on the methodology of the study, which explains how the data were collected. The chapter identified the research design, which connects to psychometric theory. The use of SDA means the data analysis will be limited, as the data have already been collected.

The ethical considerations relating to this study as well as the ethical clearances obtained for the data collection and study were discussed. This discussion acknowledged the need to be ethical and adhere to the requirements for research studies.

The remaining section of the chapter discussed the various data analysis techniques involved in the study. This included the reasons for selecting these methods as well as the advantages thereof.

These data analysis techniques included the descriptive statistics, Rasch analyses, CFA, MTMM analyses, DTF, and reliability analyses. The process in which these different techniques would be conducted and interpreted was also mentioned. As a result, this chapter provides an introduction to the presentation of findings.

CHAPTER 7: PRESENTATION OF RESULTS

7.1 Introduction

The previous chapter outlined the procedures that were followed to analyse the data, which will be presented in this chapter. This chapter will display the results for both ECT versions 1.2 and 1.3. The results that will be presented in this chapter are descriptive statistics, Rasch analyses, confirmatory factor analyses (CFA), multi-trait multi-method (MTMM) analyses, differential test functioning (DTF), and reliability analyses. This chapter will present the findings of these various analyses for each version, and the discussion of these results will be done in the subsequent chapter.

7.2 Descriptive Statistics

The importance of defining the sample lies in understanding the characteristics that pertain to the sample. These characteristics are the typical features that allow one to identify the sample being used in the study. In this study, the characteristics that will be explored are gender, race, age, provinces, and the distribution of language groups for both test versions. The different home languages, first languages at school, and additional languages at school will also be explored. This section will also include the distribution of the data, which includes tests of normality. The description of the psychometric tests to be used for comparison will also be presented.

7.2.1 Description of the Sample for the ECT Version 1.2

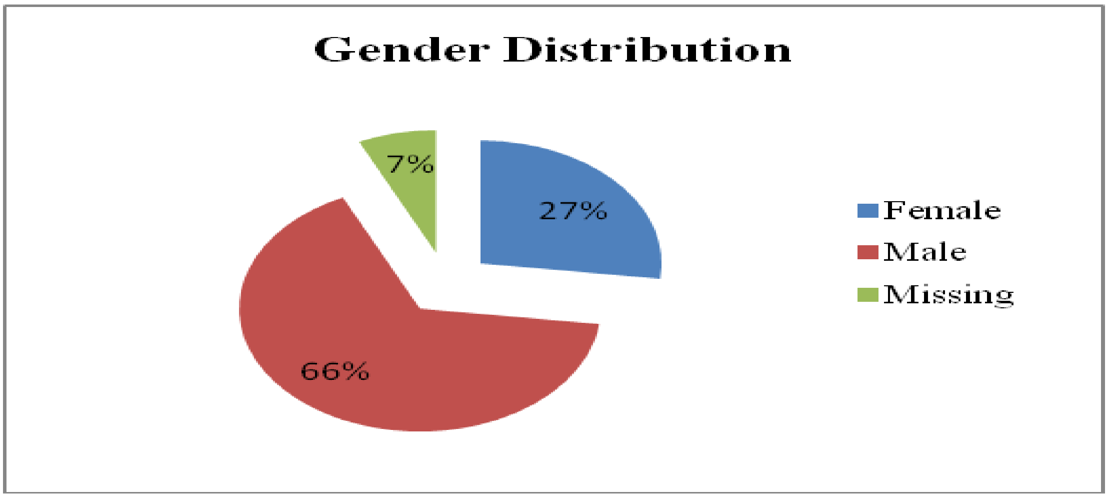


Figure 2: Graph of the Gender Distribution for ECT version 1.2

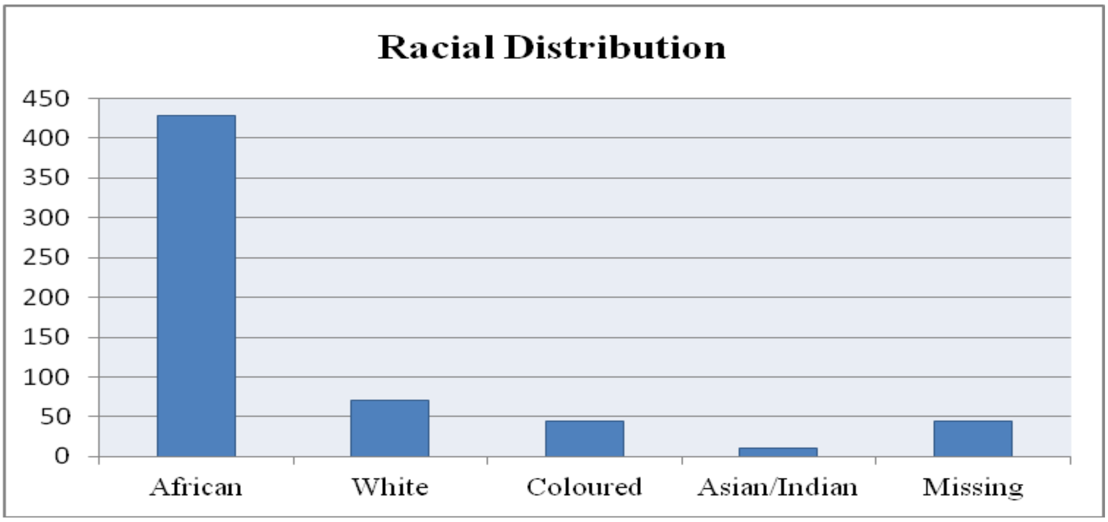


Figure 3: Graph of the Racial Distribution for the ECT version 1.2

According to the pie graph (Figure 2), the gender breakdown reveals that males constituted the majority of the sample (n = 395), while less than a quarter were females (n = 158). There was only a small percentage (n = 44) of missing data.

The bar graph above (Figure 3) displays the racial distribution of the sample. From this graph, it can be observed that the majority of the individuals were African (n = 428). The remainder of the sample comprised of White (n = 71), Coloured (n = 44), and Asian/Indian (n = 10) individuals which were represented to a lesser extent. There is also a similar amount of missing data (n = 44) for the racial information.

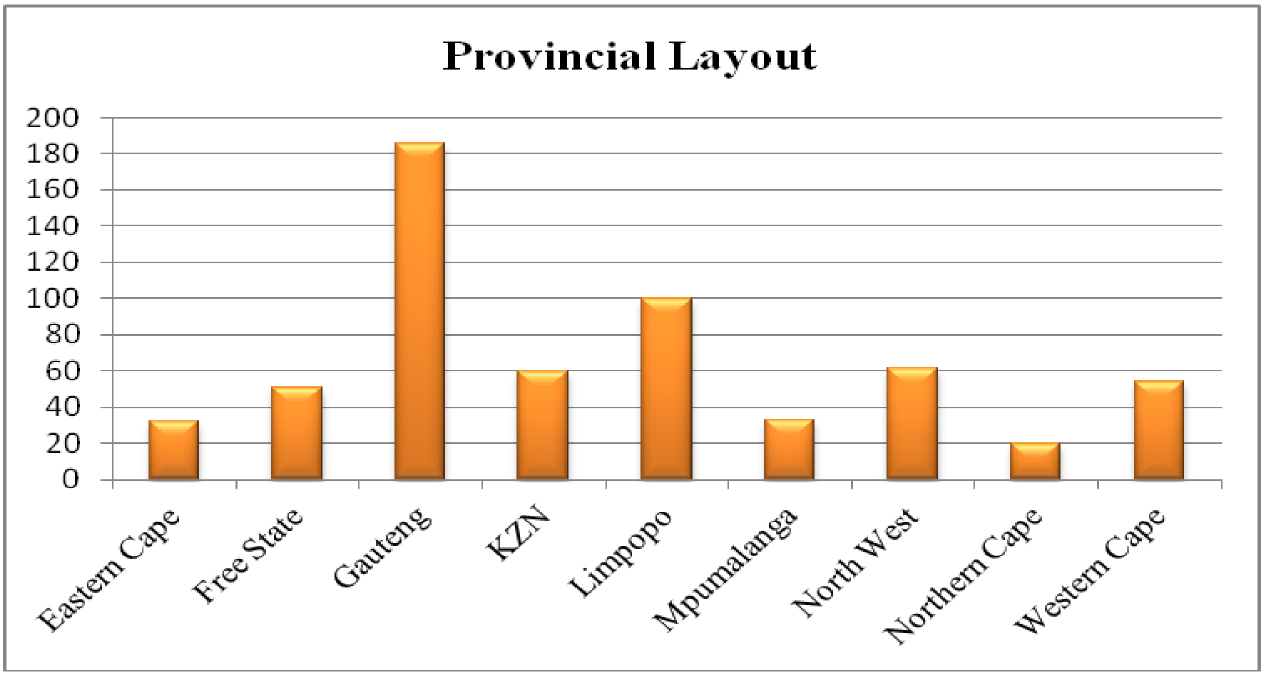


Figure 4: Graph of the Provincial Distribution for the ECT 1.2

According to the provincial distribution indicated in Figure 4 above, all nine provinces are represented by the individuals in the study. The majority of individuals resided in the Gauteng province, n = 186 (32%). The other provinces were: Limpopo, n = 100 (17%); North West, n = 61 (10%) and Kwa-Zulu Natal (KZN) province, n = 60 (10%). The least represented provinces were the Northern Cape, n = 20 (3%), Eastern Cape, n = 32 (5%), and Mpumalanga province, n = 33, (6%).

Table 2: The Distribution of Ages for the ECT version 1.2

Age	Number	Percentage
18	57	10%
19	93	16%
20	94	16%
21	77	13%
22	66	11%
23	68	11%
24	49	8%
25	29	5%
26	17	3%
27	5	1%
28	3	1%
29	2	0%
30	3	1%
31	2	0%
33	2	0%
34	4	1%
35	1	0%
36	4	1%
37	6	1%
38	3	1%
40	4	1%
41	1	0%
42	2	0%
43	1	0%
45	1	0%
50	1	0%
52	1	0%
Missing	1	0%
Total	597	100%

All the ages for the sample are listed in Table 2 above. From this table, it is evident that the majority of the individuals were either 19 or 20 years old. The youngest age group was 18, and the oldest individual was 52 years old. The mean age for this sample was 22 years.

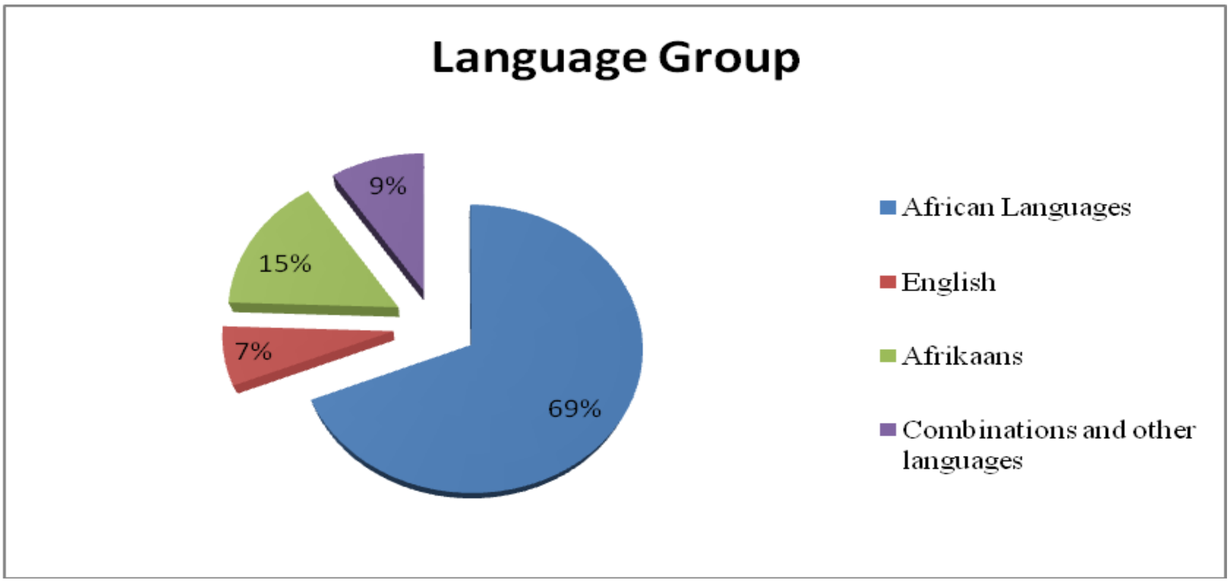


Figure 5: Graph of the Language Groupings for the ECT version 1.2

In the pie chart above (Figure 5), the home languages are grouped according to the following: African languages, English, Afrikaans, and combinations and other languages. This provides an idea of the language spread of the individuals in this sample. According to the figure, the vast majority (69%) of the candidates spoke African languages (n = 411), while 15% (n = 91) were Afrikaans language speakers. The amount of those indicating combinations of languages and other foreign languages (n = 55) was more than those indicating that they were English language speakers (n = 40).

Table 3: The Distribution of Home Language for the ECT version 1.2

Home Languages	Number	Percentage
Afrikaans	91	15%
English	40	7%
IsiNdebele	8	1%
IsiXhosa	39	7%
Sepedi	82	14%
Sesotho	40	7%
SiSwati	14	2%
Setswana	108	18%
Tshivenda	33	6%
Xitsonga	27	5%
Zulu	60	10%
Bulgarian	1	0%
Zulu/IsiXhosa	2	0%
Afrikaans/German	1	0%
Afrikaans/Setswana	2	0%
English/Afrikaans	15	3%
English/IsiXhosa	1	0%
English/Setswana	2	0%
English/Zulu	1	0%

English/Zulu	2	0%
Sepedi/Afrikaans	1	0%
Sepedi/IsiNdebele	1	0%
Sepedi/Sepedi	1	0%
Sepedi/Tshivenda	1	0%
Sepedi/Zulu	3	1%
Sesotho/Afrikaans	2	0%
Sesotho/English	2	0%
Sesotho/English/Zulu	1	0%
Sesotho/IsiXhosa	1	0%
Sesotho/Setswana	1	0%
Sesotho/Zulu	3	1%
SiSwati/English	1	0%
Setswana/Afrikaans	1	0%
Setswana/Sepedi	1	0%
Setswana/Sesotho	1	0%
Setswana/Xitsonga	3	1%
Setswana/Zulu	1	0%
Xitsonga/Sepedi	1	0%
Tsonga, Zulu	1	0%
Total	596	100%

Within Table 3, there is also an individual who indicated Bulgarian as his or her home language, which could be due him or her being a foreigner who immigrated to South Africa. Another observation was the combination of Afrikaans and German specified by one individual, who may also have been a foreigner or his or her family might have immigrated and kept German within their home setting.

There are many combinations of languages used as home languages, but the majority of these are English and Afrikaans (3%). There were 14 different combinations of African languages indicated, and in one case, three languages were spoken within the home environment.

Table 4: The Distribution of First Languages at School for the ECT version 1.2

First Languages	Number	Percentage
Afrikaans	88	15%
English	176	29%
IsiNdebele	4	1%
IsiXhosa	23	4%
Sepedi	71	12%
Sesotho	26	4%
SiSwati	8	1%
Setswana	101	17%
Tshivenda	25	4%
Xitsonga	12	2%
Zulu	47	8%
English and Xitsonga	1	0%
English/Afrikaans	6	1%
English/Afrikaans/Setswana	1	0%
English/Setswana	3	1%
English/Tshivenda	1	0%
Sesotho/English	2	0%

IsiXhosa/Afrikaans	1	0%
Tshivenda/Sesotho	1	0%
Total	597	100%

Table 4 above represents the first language choices of individuals at school. This refers to the language they communicate with as their first language within their school setting as well as the language that would be indicated as their first language on their Grade 12 results. In this table, it is clear that all 11 languages are represented by the individuals in the study. There are also a few individuals who have indicated two first languages subjects at school. This has not been validated in any way, but it is possible at some schools to have two languages as first language subjects.

According to the distribution of the languages (Table 4), the majority of the individuals have English as their first language, which is an expected norm for schools in South Africa. There are also a substantial amount of the individuals who have the following languages as first languages: Setswana (17%), Afrikaans (15%), and Sepedi (12%). The smallest first language groups identified out of the 11 official languages were IsiNdebele (1%), SiSwati (1%), and Xitsonga (2%).

Table 5: The Distribution of Additional Languages at School for the ECT version 1.2

Additional Languages	Number	Percentage
Afrikaans	125	21%
English	355	59%
IsiXhosa	6	1%
Sepedi	9	2%
Sesotho	7	1%
SiSwati	2	0%
Setswana	12	2%
Sotho	1	0%
Tshivenda	3	1%
Xitsonga	1	0%
Zulu	8	1%
Afrikaans/Zulu	2	0%
English/Afrikaans	49	8%
Sepedi/Afrikaans	1	0%
Sesotho/English	1	0%
N/A	12	2%
Missing	3	1%
Total		100%

Table 5 shows that all 11 languages are represented as second languages and a few indicated dual second languages. The indication of two second languages on a Grade 12 certificate is possible. This, however, only occurs when there is also a home or first language specified, which would, in essence, mean that the individual has three languages indicated on their Grade 12 certificate (one first language and two additional languages). The majority of the individuals indicated that English (59%) was their additional language, while numerous individuals reported Afrikaans as an additional language. The other languages were reported by relatively few individuals as additional languages. There are, however, a relative number of individuals (8%) who stated that English and Afrikaans are their additional languages. This would most probably be the case for the individuals who have an African language as their first language. Within this table, some individuals indicated N/A (Not Applicable), which would suggest that they have two first languages, as all individuals need to have at least two languages on their Grade 12 certificate. There are only a few cases of missing data.

7.2.2 Description of the Sample for the ECT Version 1.3

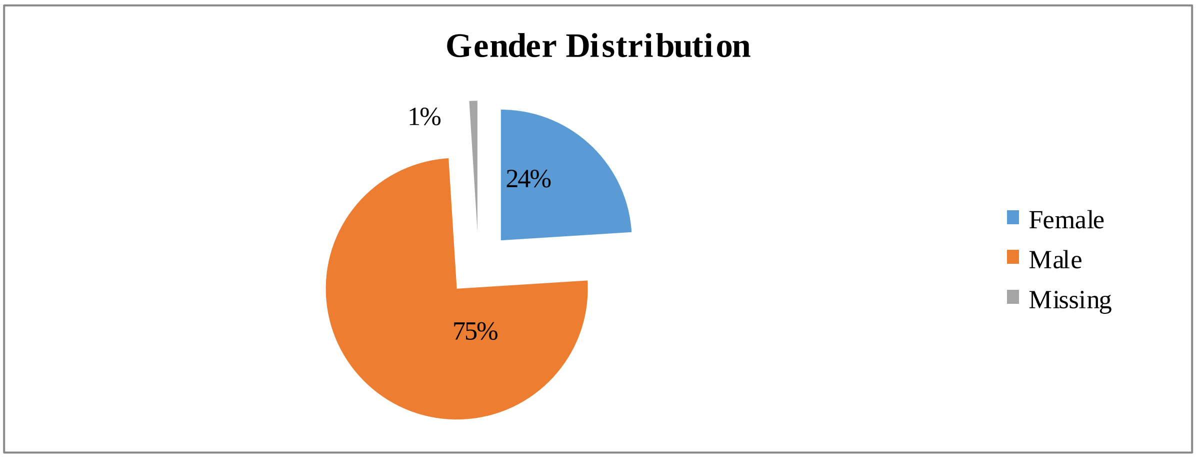


Figure 6: Graph of the Gender Distribution for the ECT version 1.3

The pie chart above (Figure 6) indicates the gender distribution of the sample of the ECT version 1.3. The majority of the individuals were males (n = 666), with less than a quarter females (n = 212). There was also very few cases of missing data (n = 4).

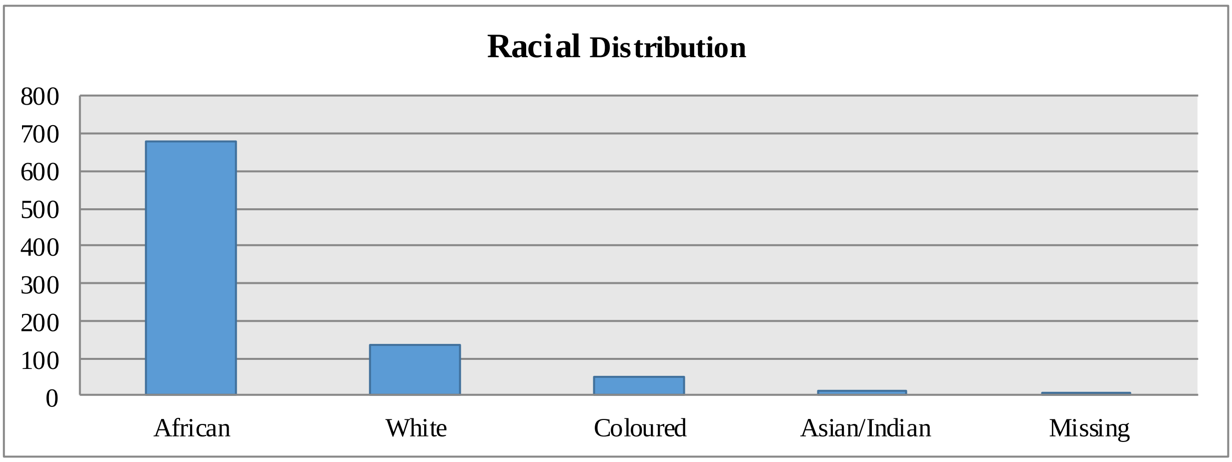


Figure 7: Graph of the Racial Distribution for the ECT version 1.3

According to the racial distribution (Figure 7), the majority of the individuals were African (n = 682). The White group (n = 135) was relatively larger than the Coloured group (n = 50). The smallest grouping was the Asian/Indian group (n = 11). There were also very few cases of missing data (n = 4).

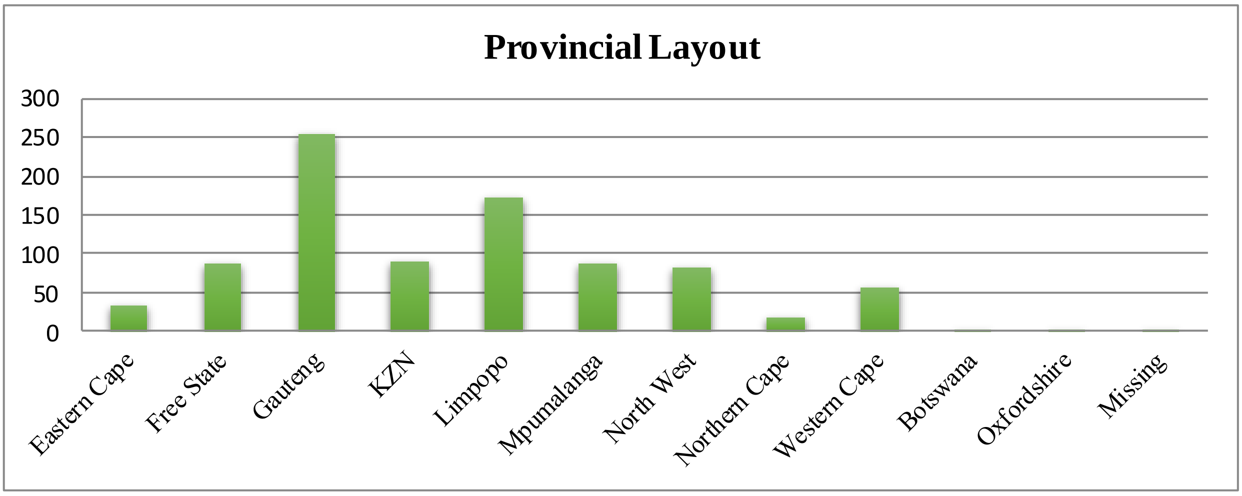


Figure 8: Graph of the Provincial Distribution for the ECT version 1.3

In Figure 8, the distribution of the provinces for the sample is displayed. From this figure, the majority of the individuals reside in the Gauteng (n = 253) and Limpopo (n = 171) provinces. There are also a considerable number of individuals from KZN (n = 90), the Free

State (n = 88), Mpumalanga (n = 88), and North West province (n = 83). The least number of individuals reside in the Western Cape (n = 55), Eastern Cape (n = 33), and Northern Cape (n = 18). There were two cases indicated that were not provinces; Botswana (n = 1) and Oxfordshire (n = 1). There was only one case of missing data.

Table 6: The Distribution of Age for the ECT version 1.3

Age	Number	Percentage
18	170	19%
19	163	19%
20	152	17%
21	126	14%
22	90	10%
23	82	9%
24	43	5%
25	23	3%
26	12	1%
27	2	0%
28	2	0%
29	2	0%
30	1	0%
31	3	0%
36	2	0%
37	2	0%
41	1	0%
42	1	0%
Missing	5	1%
Total	877	100%

In Table 6 above, the different ages are listed for the sample. The youngest age is 18 years old, while the oldest is 42 years old. The majority of the individuals in the sample were 18 to 19 years old. The mean age of the sample was 21 years old.

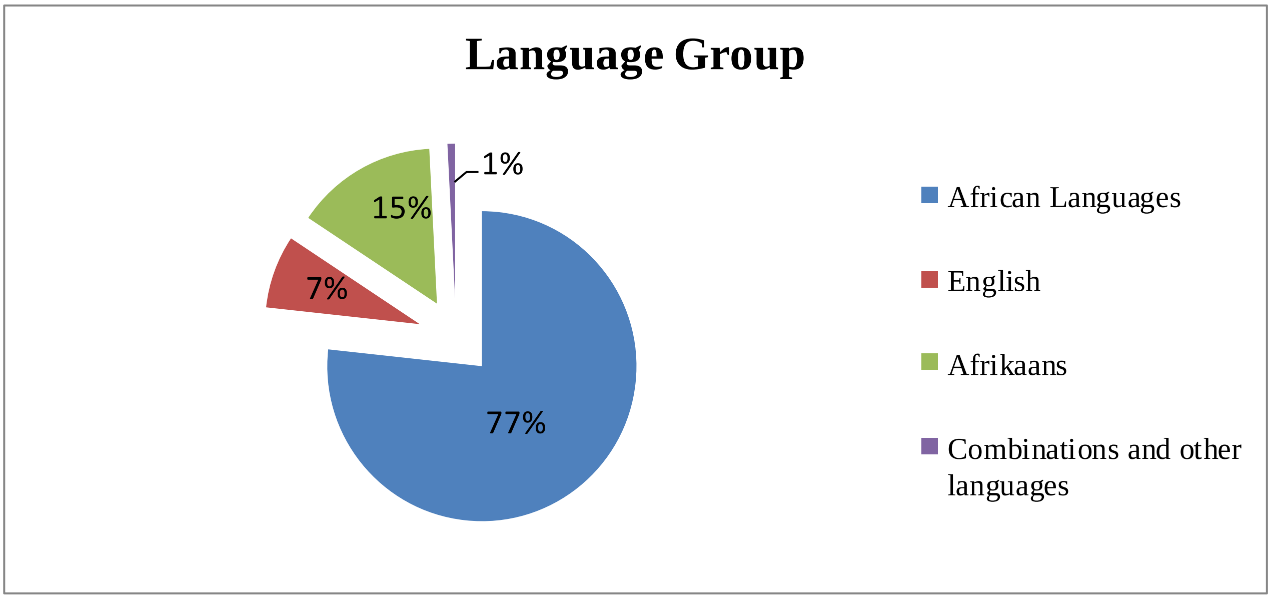


Figure 9: Graph of the Language Grouping for the ECT version 1.3

Figure 9 above displays the language groupings of the sample. These groupings are based on the individual’s home language. From this figure, it is evident that the majority of the sample were African language speakers (n = 676). There was a considerable number of individuals who were Afrikaans language speakers (n = 131), while the smallest number was English language speakers (n = 67). Very few individuals indicated a combination of languages and other foreign languages (n = 7).

Table 7: The Distribution of Home Language for the ECT version 1.3

Home Languages	Number	Percentage
Afrikaans	131	15%
English	67	8%
IsiNdebele	28	3%

Sepedi	144	16%
SiSwati	27	3%
Sotho	82	9%
Tsonga	55	6%
Setswana	119	13%
Venda	57	6%
Xhosa	64	7%
Zulu	98	11%
English/ Afrikaans	3	0%
Setswana/ Afrikaans	1	0%
Portuguese	1	0%
Other	2	0%
Missing	3	0%
Total	882	100%

Table 7 above lists the home languages of the individuals for the ECT version 1.3. According to this table, the majority of the individuals spoke Sepedi, Afrikaans, and Setswana in their homes. The least spoken languages are isiNdebele and SiSwati. All 11 languages are, however, represented in this table, as well as combinations of languages. There are only a few individuals who indicated a combination of languages as their home language, and there was one foreign language, Portuguese, reported. It is not clear what "other" indicates, and only a few missing responses were observed.

Table 8: The Distribution of First Language for the ECT version 1.3

First Languages	Number	Percentage
Afrikaans	117	13%
English	272	31%
IsiNdebele	8	1%
Sepedi	111	13%
SiSwati	19	2%
Sotho	50	6%
Tsonga	30	3%
Setswana	110	12%
Venda	43	5%
Xhosa	36	4%
Zulu	74	8%
English/ Afrikaans	1	0%
English/ Setswana	1	0%
Sepedi/ Zulu	1	0%
Missing	9	1%
Total	882	100%

Table 8 represents the first languages the individuals had as subjects at school. From this table, the majority of the individuals had English as their first language. There is also quite a number who had Afrikaans, Sepedi, and Setswana as first languages. Although all 11 languages are represented in this table, the least represented languages were isiNdebele, SiSwati, and Tsonga. There were combinations of languages indicated by very few individuals and a small number of missing data were observed.

Table 9: The Distribution of Additional Languages for the ECT version 1.3

Additional Languages	Number	Percentage
Afrikaans	154	17%
English	571	65%
Sepedi	14	2%
SiSwati	1	0%
Sotho	9	1%
Tsonga	4	0%
Setswana	22	2%
Venda	7	1%
Xhosa	2	0%
Zulu	16	2%
German	2	0%
English/ Afrikaans	33	4%
English/ Xhosa	1	0%
Afrikaans/ Sepedi	1	0%
Afrikaans/ Sotho	1	0%
Sepedi/Tsonga	1	0%
Sotho/ Setswana	1	0%
Other	1	0%
Missing	40	5%
Total	882	100%

Table 9 above lists the second or additional language subject that individuals had at school. According to the table, the vast majority of individuals had English as their second or additional language. There were also numerous individuals who had Afrikaans as a second language at school. Only 10 official languages are present in the table because isiNdebele was not selected by any participants. There was one foreign language, German, which was reported by two individuals. There were combinations of languages indicated by a few individuals, but most were the combination of English and Afrikaans. There was also a number of missing data.

7.2.3 Description of the Data for ECT Version 1.2

Tables 10 and 11 below indicate the distribution of the data for the ECT version 1.2. Several descriptive statistics provide information about the distribution of the data for the sample. According to the mean (23), median (24), and trimmed mean (23) values, there is a similar trend for how individuals performed in the test. These values suggest the individuals were able to answer a large portion of the test correctly, since the total for the test is 39. This means that, on average, individuals got 59% of the test content correct. The minimum (8) and maximum (38) values signify that no individual got less than 21% of the test content correct, while the most test content correctly covered by an individual was 97%. This suggests that the test could have been experienced as relatively easy.

Table 10: The Distribution of the Data for the ECT version 1.2

Descriptives	Statistical values
Mean	23.51
5% Trimmed Mean	23.56
Median	24.00
Variance	31.643
Std. Deviation	5.585
Minimum	8
Maximum	38
Range	30
Interquartile Range	8
Skewness	-.125
Kurtosis	-.284

Table 11: Tests for Normality for the ECT version 1.2

Tests of Normality	Statistics	Degrees of Freedom	Significance
Kolmogorov-Smirnov	.052	597	.001
Shapiro-Wilk	.994	597	.019

The normality of the data is determined by reviewing several statistics such as the skewness, kurtosis, Kolmogorov-Smirnov, and Shapiro-Wilk tests. These statistics denote whether the data of the sample are normally distributed. According to the values for the skewness (-0.125) and kurtosis (-0.284), the data are negatively skewed and have a flat distribution. This suggests that the data are not normally distributed. The Kolmogorov-Smirnov and Shapiro-Wilk tests (Table 11) are used to assess normality. According to the values for the Kolmogorov-Smirnov and Shapiro-Wilks tests respectively, $D(597) = 0.052$, $p < 0.001$ and $D(597) = 0.994$, $p < 0.05$, the data are significantly non-normal. One can conclude from these various statistics that the data are not normally distributed for this sample.

7.2.4 Description of the Data for ECT Version 1.3

The description of the data is explored by examining statistics that provide information about the sample data. In Table 12 and 13, the mean (26), median (26), and trimmed mean (26) are the same value, suggesting there was similarity in how the individuals performed in the test. This also indicates that the individuals were able to answer a relative portion of the test correctly. According to the mean value, the average performance of individuals in this test was 60%, as they obtained an average of 26 out of 42. The minimum (8) and maximum (39) values obtained mean the lowest percentage attained was 19%, while the highest percentage was 93%. There was a wide range between the highest and lowest score.

Table 12: The Description of the Data for the ECT version 1.3

Descriptives	Statistical values
Mean	25.92
5% Trimmed Mean	26.02
Median	26
Variance	30.511
Std. Deviation	5.524
Minimum	8
Maximum	39
Range	31
Interquartile Range	8
Skewness	-0.256
Kurtosis	-0.140

Table 13: Tests for Normality for the ECT version 1.3

Tests of Normality	Statistics	Degrees of Freedom	Significance
Kolmogorov-Smirnov	.053	881	.000
Shapiro-Wilk	.991	881	.000

According to the skewness (-0.256) and kurtosis (-0.140) values, the data are negatively skewed and have a flat distribution. The data are consequently not normally distributed. The tests of normality are, however, also considered before determining normality. The Kolmogorov-Smirnov and Shapiro-Wilk tests (Table 13) indicate the following respectively: $D(881) = 0.053$, $p < 0.001$ and $D(881) = 0.991$, $p < 0.001$. This indicates that the data are significantly non-normal. Thus, one can conclude that the data are not normally distributed.

7.2.5 Description of the Psychometric Tests With the ECT Versions 1.2 and 1.3

Several psychometric tests will be used for the correlation analysis to which the ECT must be compared. For this reason, it is essential to explore the descriptive statistics of these psychometric tests, which involves assessing the mean values and standard deviations. It should be noted that the sample sizes are the same for the psychometric tests for the respective test versions, as they were collected in groupings.

Table 14: Descriptives for the Academic Aptitude Tests with ECT version 1.2

Descriptive Statistics				
Tests	Total Test Score	Mean	Std. Deviation	Sample size
ECT 1.2	39	23.76	5.361	272
AAT 1	31	18.07	6.857	272
AAT 2	30	14.82	5.508	272
AAT 3	30	12.17	5.843	272
AAT 4	30	14.04	6.093	272
AAT 5	30	9.02	5.135	272

When reviewing the mean values for the different AAT tests and the ECT version 1.2 (Table 14), the most apparent observation is that the ECT version 1.2 has the highest mean value (24). The mean value indicates that, on average, individuals obtained 62% for the ECT. This means that individuals scored far better on the ECT than on the other measures.

The average performance on the AAT tests was as follows: AAT 1: an average of 58%; AAT 2: an average of 50%; AAT 3: an average of 40%; AAT 4: an average of 47%; and AAT 5: an average of 30%. It is evident from these averages that individuals performed relatively poorer on these tests, especially the AAT 3 and AAT, than the ECT. This suggests that the AAT tests have a higher difficulty level than the ECT and thus there is a significant difference in performance of individuals in these different psychometric tests. The standard deviations for the AAT tests and the ECT version 1.2 are all different from each other, which indicate that individuals performed differently on the various tests. This serves to endorse the percentages of the mean performances of individuals on the different psychometric tests.

Table 15: Descriptives for the Differential Aptitude Tests and Senior Aptitude Tests (ECT version 1.2)

Descriptive Statistics				
Tests	Total Test Score	Mean	Std. Deviation	Sample size
ECT 1.2	39	23.41	5.847	272
DAT 2	40	13.79	4.287	272
DAT 3	25	15.93	4.611	272
DAT 9	25	9.70	4.220	272
DAT 10	25	12.11	4.465	272
SAT 2	25	18.35	8.098	271
SAT 4	30	16.88	3.792	272
SAT 5	30	13.64	4.915	272
SAT 6	30	16.11	6.478	272
SAT 7	30	16.72	7.003	272
SAT 8	30	15.24	5.436	272
SAT 10	30	19.87	5.885	272

The tests in Table 15 were all administered to the same group of individuals. The ECT has the second highest test total and the highest mean compared to the other psychometric tests. The mean score of ECT version 1.2 is furthermore much higher than all the DAT and SAT test means. This mean value (23) denotes that individuals obtained an average of 59% for the ECT version 1.2. This would suggest that individuals performed above average on the test.

The test performance of individuals on these different psychometric tests can be shown as follows: DAT 2: an average of 35%; DAT 3: an average of 64%; DAT 9: an average of 40%; DAT 10: an average of 48%; SAT 2: an average of 72%; SAT 4: an average of 57%; SAT 5: an average of 47%; SAT 6: an average of 53%; SAT 7: an average of 57%; SAT 8: an average of 50%; and SAT 10: an average of 67%. When reviewing the performance of individuals on these tests, it is apparent that tests such as the SAT 2, SAT 10, and DAT 3 are considered easier than the others. The more difficult tests are the DAT 2, DAT 9, SAT 5, and DAT 10. The other DAT, SAT, and ECT tests fall within a mediocre range and produced just above average performance. The standard deviation scores for the psychometric tests differ significantly from each other, and this suggests that the individuals performed very differently on these tests. This confirms the percentages of the test performance of individuals in the various psychometric tests.

Table 16: Descriptives for the Academic Aptitude Tests with ECT version 1.3

Descriptive Statistics				
Tests	Total Test Score	Mean	Std. Deviation	Sample Size
ECT 1.3	42	26.36	4.901	211
AAT 1	31	17.40	5.466	211
AAT 2	30	14.63	4.564	211
AAT 3	30	11.50	5.301	211
AAT 4	30	13.16	5.300	211
AAT 5	30	8.86	4.628	210

The ECT version 1.3 has both the highest total test score and mean score compared to the AAT tests. The test performance on the ECT version 1.3 indicates that individuals obtained 62%, on average. This implies that individuals found the test relatively easy. The test performance on the AAT tests was as follows (Table 16): AAT 1: an average of 55%; AAT 2: an average of 50%; AAT 3: an average of 40%; AAT 4: an average of 43%; and AAT 5: an average of 30%. This shows that individuals performed much better on the ECT than the AAT tests. The AAT 5, AAT 3, and AAT 4 were experienced as quite difficult for these individuals. This deduction is supported by observing the standard deviation scores for the various psychometric tests.

Table 17: Descriptives for the Differential Aptitude Tests and Senior Aptitude Tests (ECT version 1.3)

Descriptive Statistics				
Tests	Total Test Score	Mean	Std. Deviation	Sample Size
ECT 1.3	42	25.54	5.598	648
DAT 2	40	12.58	3.945	647
DAT 3	25	15.64	4.413	647
DAT 9	25	9.01	3.859	647
DAT 10	25	10.85	4.346	647
SAT 2	25	16.93	7.666	646
SAT 4	30	15.67	3.735	646
SAT 5	30	12.88	4.810	646
SAT 6	30	15.40	6.104	646
SAT 7	30	15.63	6.942	647
SAT 8	30	15.14	5.501	647
SAT 10	30	18.74	5.963	647

From the above table, it is apparent that the ECT version 1.3 total test score and mean score were the highest in both instances. In terms of the test performance on the ECT version 1.3 for this group, individuals obtained an average of 60%. This suggests that they were able to correctly complete the majority of the ECT, indicating that the test was not difficult for individuals. In terms of the test performance for the DAT and SAT tests (Table 17), the following was found: DAT 2: an average of 33%; DAT 3: an average of 64%; DAT 9: an average of 36%; DAT 10: an average of 44%; SAT 2: an average of 68%; SAT 4: an average of 53%; SAT 5: an average of 43%; SAT 6: an average of 50%; SAT 7: an average of 53%; SAT 8: an average of 50%; and SAT 10: an average of 60%. Based on these percentages, it is evident that individuals experienced the DAT 2, DAT 9, DAT 10, and SAT 5 as quite difficult.

The remaining DAT, SAT, and ECT tests fall within average to above average performance. When observing the standard deviation scores for the different psychometric tests, the differences in performance for these tests are confirmed. The standard deviation scores for these tests differ quite dramatically, showing that individuals did not perform similarly across the tests.

7.3 Rasch Analyses Results

The Rasch analyses results were obtained by following the procedures listed in Chapter 6. The output for the Rasch analyses will be divided according the two test versions. The ECT version 1.2 was the initial test version and consisted of multiple choice questions for both the comprehension and the language section; and written answers for the sentence construction section. The sections of this test version were not clearly demarcated and the instructions were minimal. There was a time limit of 45 minutes imposed and the test consisted of 39 items.

7.3.1 ECT Version 1.2 Results

The output will be presented in the following order: the fit statistics (the person statistics, the item statistics, and test empirical randomness); the summary of category structure statistics; the person-item map; the measure order statistics (the bubble chart); the misfit order statistics; dimensionality (the variance decomposition of observations, the standardised residual contrast plots, the standardised residual loadings); and characteristic curves (the test characteristic curve, the item characteristic curve and the test information function curve).

7.3.1.1 Fit statistics

The fit statistics consist of the person statistics and the item statistics. The results will inform one whether the items fit the Rasch model and whether the persons are responding in an expected way. This will also include the test empirical randomness, which indicates whether the test is severely influenced by random responses.

7.3.1.1.1 The person statistics

Tables 18 and 19 below indicate the model fit information. The average person statistics are presented in Table 18. The mean is 23.5, which corresponds to the mean indicated in the descriptive statistics. The fit statistics (infit and outfit MNSQ) need to be between 0.7 and 1.2 (Linacre, 2002; Smith, Schumacker, & Bush, 1998) or 1.3 (Bond & Fox, 2007; Linacre, 2011; Pensavalle & Solinas, 2013). The average infit and outfit MNSQ values are 1.01, which is considered good and indicates that the person's abilities fit the model on average. The average ZSTD for both the infit and outfit are well within range (not statistically significant misfit) and show that the persons fit the model. The person separation value of 1.87 is small (Baghaei & Amrahi, 2011) and indicates that there is not much variation among the abilities of the persons. This means that the persons in the sample mostly have the same ability. The candidate reliability is .78, which is not a very good reliability value in terms of classical test theory (CTT) (Linacre, 2012c).

The irregular patterns of the minimum and maximum values for the infit and outfit MNSQ indicate that there are, however, persons who do not fit the model. This is evident by the minimum values (0.59 and 0.43) and maximum values (1.60 and 9.56). These maximum and minimum values show that there are problems with the response patterns of the persons in the model and thus it needs to be investigated. The maximum (2.8 and 3.3) and minimum (-2.6 and -1.9) ZSTD values are cause for concern as they exceed 1.96 (Linacre, 2011, 2012b, 2012c), which indicates that there is a statistically significant deviation from the model. The ZSTD should correspond to MNSQ. The standard error of candidate mean is 0.04, which can be considered low and indicative of less error in the measurement of the candidate mean. The standard error of measurement is an average measurement error for the persons of the test and is based on CTT. The standard error of measurement value for the persons of the test is 2.59, which indicates that there is a small amount of error in the average measurement of a person in the model. (Linacre, 2012c; 2016).

Table 18: Average Person Statistics for ECT version 1.2

	Total Score	**Measure**	**MNSQ Infit**	**MNSQ Outfit**	**ZSTD Infit**	**ZSTD Outfit**
Mean	23.5	.61	1.01	1.01	.0	.0
Max	38.0	4.44	1.60	9.56	2.8	3.3
Min	8.0	-1.79	.59	.43	-2.6	-1.9

7.3.1.1.2 The item statistics

The average item statistics are indicated in Table 19. The average infit and outfit MNSQ values were 0.99 and 1.01 respectively, which are considered good. They indicate that, on average, the items fit the model. When observing the infit and outfit MNSQ maximum values (1.29 and 1.79) and minimum values (0.81 and 0.58), it becomes apparent that there are some items which are not functioning as expected and do not fit the model. These values therefore indicate that there are unexpected patterns within the items (since they deviate from the model) that need to be observed, these values are however not particularly extreme. Since the infit MNSQ is mostly within range, the concern is that there is possible noise or outliers present due to the irregularity in the outfit MNSQ values (Linacre, 2011; Maree, 2004b, 2004c).

The ZSTD for both the mean infit and outfit are not statistically significant, which indicates that the items are not misfitting. The maximum (8.6 and 9.2) and minimum (-4.7 and -3.8) infit and outfit ZSTD values are very large and are thus statistically significant. This implies that there are items that are extremely misfitting in the data. The item separation index is 11.93, which is rather large and shows that there is a broad distribution of item difficulties. This would, however, be observed when viewing the item-person map. The standard error of test mean is 0.22, which is very small, suggesting that there is little error. In addition to this, the item reliability is .99 which is regarded as excellent (Linacre, 2011, 2012b, 2012c).

Table 19: Item Statistics for ECT version 1.2

	Measure	**MNSQ Infit**	**MNSQ Outfit**	**ZSTD Infit**	**ZSTD Outfit**
Mean	.00	.99	1.01	.0	.2
Max	2.78	1.29	1.79	8..6	9.2
Min	-3.04	.81	.58	-4.7	-3.8

7.3.1.1.3 Test empirical randomness

The empirical randomness for the test is shown in Figure 10 below. From this figure, one can see the deviation from the expected randomness in the test through the variance of the MNSQ outfit and infit (Linacre, 2015, pp. 458-459). One can see that the MNSQ outfit is the greatest contributor to the test randomness, while the MNSQ infit deviates slightly. The MNSQ outfit is larger towards the higher end of the measure of the latent variable.

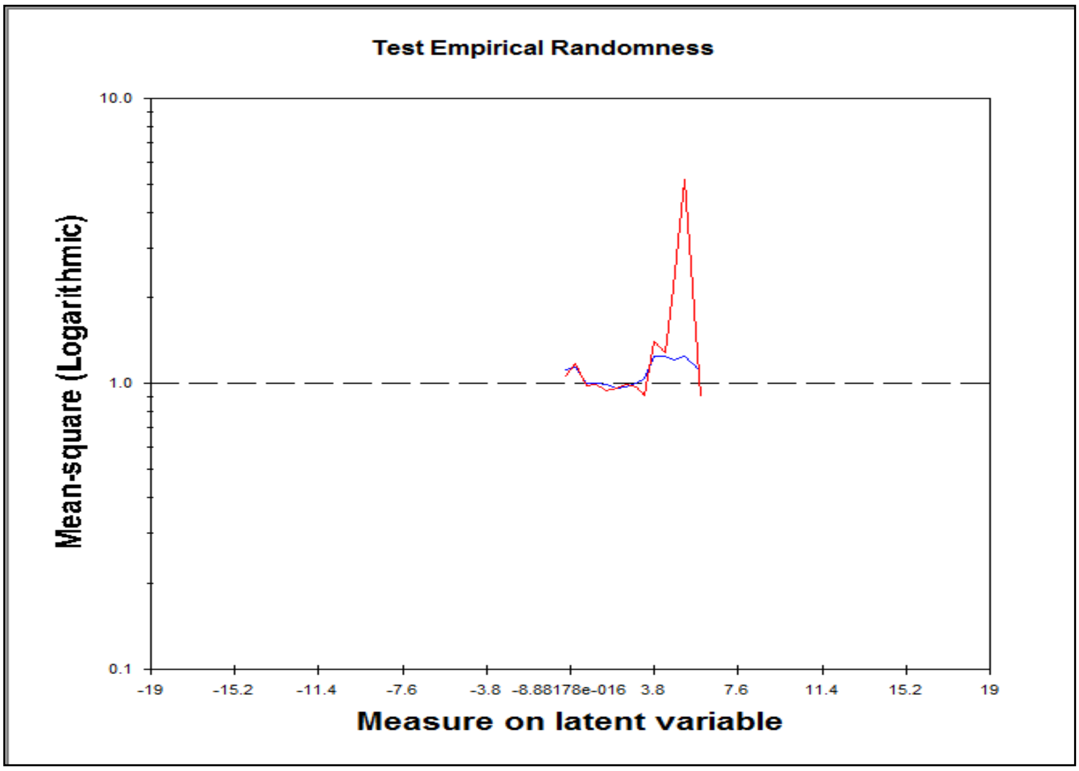


Figure 10: Test Empirical Randomness of the ECT version 1.2

7.3.1.1.4 Summary of category structure statistics

The summary of the category structure shown in Table 20 below presents the values attributed to each of the categories. There are two categories identified, since it is a dichotomous scale. Category 1 is for correct responses and category 0 indicates incorrect responses. It should be noted that this table contains the averages for the test and for this reason, the individual items may deviate. Category 0 comprised 40% (9247) of the total responses for the test, while category 1 comprised 60% (14036) of the total responses.

Table 20: Summary of Category Structure Statistics for ECT version 1.2

Category label	Observed Average	Expected Average	MNSQ Infit	MNSQ Outfit
0	-.46	-.46	1.00	1.04
1	1.32	1.32	1.00	.97

Responses by candidates fell into both categories, although more responses fell into category 1 than category 0. This means that most candidates were able to answer the questions in the test correctly. The average measure, which is expressed as logits, increased vastly from -0.46 to 1.32. Furthermore, the observation average values are equal to the expected values. The infit MNSQ for both categories (1 and 0) was 1.00, which is an acceptable value. The outfit MNSQ values for both categories are within range of 1. Category 1 = 0.97 and category 0 = 1.04, which are acceptable, outfit values. This indicates that there is no noise or unexpected observations present in the responses for the test. This table is, however, more informative when used with rating scales (where there are more than two options to choose for the test) (Maree, 2004b).

The probability curve of observation for the two response categories of the ECT can be seen in Figure 11. Category 0 ranges from just below 0.9 to 0.10, while category 1 ranges from 0.10 to 0.90. In this figure, one can see that there is an increase in incorrect answers in low ability candidates and a decrease in incorrect answers for high ability candidates. The probability of answering correctly increases, while the probability of answering incorrect decreases. This graph is more informative when used to display rating scales patterns (Linacre, 2011; Maree, 2004b).

```
        DICHOTOMOUS CURVES
P      -+--------------+--------------+--------------+--------------+-
R  1.0 +                                                             +
O      |                                                             |
B      |0                                                           1|
A      | 000000                                               111111 |
B   .8 +       00000                                     11111       +
I      |            0000                             1111            |
L      |                0000                     1111                |
I      |                    000               111                    |
T   .6 +                       000         111                       +
Y      |                          000   111                          |
    .5 +                             ***                             +
O      |                          111   000                          |
F   .4 +                       111         000                       +
       |                    111               000                    |
R      |                1111                     0000                |
E      |            1111                             0000            |
S   .2 +       11111                                     00000       +
P      | 111111                                               000000 |
O      |1                                                           0|
N      |                                                             |
S   .0 +                                                             +
E      -+--------------+--------------+--------------+--------------+-
       -2             -1              0              1              2
       Candidate [MINUS] ECT MEASURE
```

Figure 11: Probability Curves of Observations in the two Categories of the ECT version 1.2

7.3.1.2 Person-item map

Since there seem to be some discrepancies between the item difficulty and person's ability, the person-item map will assist in visually demonstrating how these items fit in comparison to the person's ability. The map is shown below (Figure 12), with dotted rectangles to illustrate one of the important aspects of the map which will be discussed below.

```
MEASURE Candidate - MAP - ECT
               <more>|<rare>
         5                +
                           |
                           |
                        .  |
                           |
                           |
          4                +
                           |
                        .  |
                           |
                           |
                        .  |
          3                +
                        .  |  I0037
                           |T
                       .#  |  I0036
                          T|  I0021  I0039
                       ##  |
          2           .##  +
                      ###  |  I0023
                     .###  |
                          S|  I0001
                    .####  |S
                   .#####  |  I0022
          1 .############  +  I0005  I0038
                  .######  |  I0009  I0016
                 ######## M|  I0014
                  #######  |
                   .#####  |  I0030
                   .#####  |  I0007  I0015  I0019  I0026  I0032
          0    .#########  +M I0011  I0020  I0024
                    #####  |  I0027  I0028
                    .#### S|  I0002  I0004  I0006  I0033
                      .##  |  I0003  I0018
                      .##  |  I0029  I0031
                       .#  |
         -1           .##  +
                        . T|
                        .  |S I0008
                           |  I0035
                        .  |  I0025
                        .  |  I0034
         -2                +  I0017
                           |  I0012  I0013
                           |
                           |
                           |T
                           |
         -3                +  I0010

                     <less>|<freq>

       EACH "#" IS 6: EACH "." IS 1 TO 5
```

Figure 12: Person-Item Map of ECT version 1.2

When observing the map (Figure 12), it can be seen that there is an overall good fit between the persons and the model. This is evidenced by the distribution of the items and persons according to difficulty. There are, however, a few gaps in the distribution which would require items of a particular difficulty level. The items appear to have a good spread along the item and persons map, as they tend to lie along the continuum (Maree, 2004b, 2004c).

The dotted rectangles indicate that several items are measuring the same difficulty level and are thus redundant. The gaps in the distribution indicate that items are required to fill these ability levels. Thus the redundant items can be used to fill the gaps in the distribution by editing them to address the required ability levels. It is also evident that in most cases, the person's ability fits alongside the item difficulty. Thus the items cover the abilities of most of the candidates. There are, however, persons who have a higher ability than the highest item and there are a few items that fall below their ability. This means that when an individual's ability location is above an item difficulty level, then the individual has a greater than 50% probability of correctly answering the item. The majority of the candidates are between -1 and 2 standard deviations. The candidates of the ECT seem to have an average to above average ability, in the context of their performance on the ECT. The items are distributed fairly well along the continuum, and there is a range of very easy, easy, moderate, and above average ability items. The test items do not, however, include persons whose ability is beyond above average and would benefit from advanced items (Dunne, Long, Craig & Venter, 2012; Long, 2011; Linacre, 2011; Maree, 2004b, 2004c).

7.3.1.3 Measure order statistics

The measure order statistics indicated in Table 21 show the item parameter information. These statistics demonstrate the difficulty of items by indicating how many individuals were able to answer the item correctly. This table is therefore structured from difficult to easy items (Linacre, 2011; Maree, 2004b, 2004c). When observing the values in the total score column, the most difficult items are items 37, 36, and 39, as only 78, 98, and 107 individuals respectively out of 597 individuals were able to answer these items correctly. The easiest items were items 10, 13, and 12, as 576, 552, and 550 individuals respectively out of 597 answered these items correctly. Interestingly enough, there was no item that individuals were not able to answer, and not a single item that all the individuals succeeded in answering correctly. The measure column shows the item difficulty, which is indicated by logits. In this column, one can observe that there are quite a few items above the mean of 0 logits, thus indicating that higher ability level items are required, while also indicating that there are several items below the mean which require items of a lower ability level.

Table 21: Measure Order Statistics for ECT version 1.2

Item	Total Score	Total Count	Measure
37	78	597	2.78
36	98	597	2.48
39	107	597	2.36
21	110	597	2.32
23	156	597	1.81
1	191	597	1.48
22	223	597	1.20
38	252	597	.97
5	255	597	.94
16	259	597	.91
9	262	597	.89
14	300	597	.59
30	330	597	.35

7	355	597	.15
32	357	597	.14
15	359	597	.12
19	359	597	.12
26	359	597	.12
11	367	597	.06
20	368	597	.05
24	381	597	-.06
27	393	597	-.16
28	403	597	-.25
2	410	597	-.31
4	413	597	-.33
33	414	597	-.34
6	420	597	-.40
18	426	597	-.45
3	427	597	-.46
31	447	597	-.66
29	452	597	-.71
8	505	597	-1.33
35	518	597	-1.53
25	528	597	-1.69
34	536	597	-1.84
17	540	597	-1.92
12	550	597	-2.15
13	552	597	-2.20
10	576	597	-3.04

7.3.1.3.1 The bubble chart

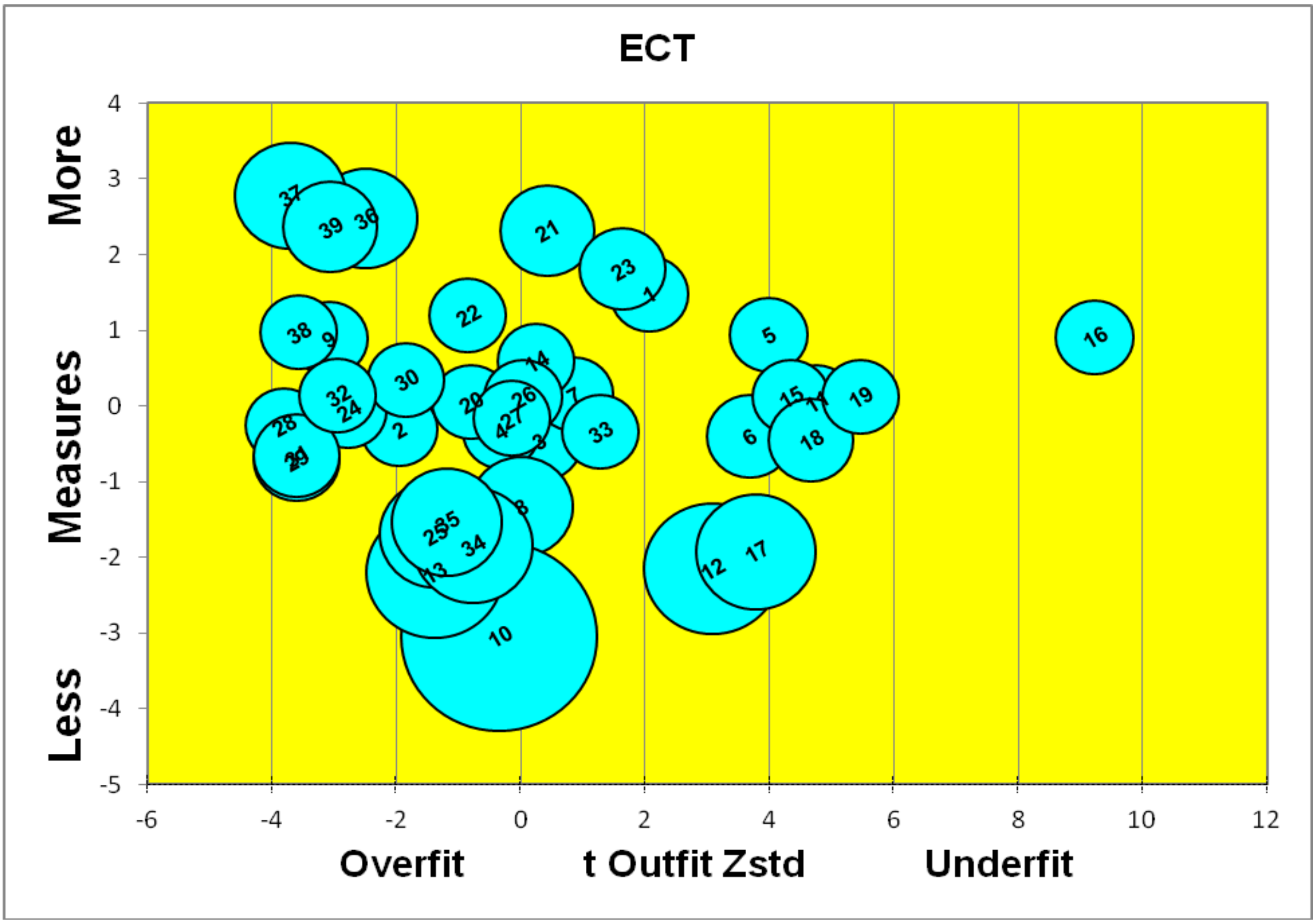


Figure 13: Bubble Chart for the Items of the ECT version 1.2

Rasch analysis, through the use of Winsteps, allows one to create a bubble chart for items (Figure 13). This is a graphical representation of the item difficulty in the form of a bubble chart. This chart visually presents the items' performance according to their difficulty (Linacre, 2015, p. 469). When viewing the chart, the items on the top, such as items 36, 37, and 39, are the most difficult items, and the items at the end, such as items 10, 12, and 13, are the easiest. This corresponds to the previous table. The size of the bubbles is significant, as it indicates that the item has either more (larger bubble) or less (smaller bubble) measurement error (Bond & Fox, 2007; Linacre, 2012a; Pae et al., 2012). One would expect the tails to have less precise estimates and thus large errors.

According to Bond and Fox (2001), it is anticipated that there would be larger errors in the measurement of items towards the extreme ends of the chart. This is observed for the easiest item (item 10), while the other extreme items previously mentioned are slightly bigger bubbles than the rest. The bubbles in the centre of the chart are relatively the same size and cluster close to each other. This would suggest that the items in the centre were more accurate, in terms of the measurement error, at measuring the candidate's ability in the ECT than the items situated on the extreme top and bottom. Although some items are not clearly visible, all the items are indicated in the bubble chart. This allows one to conclude that all the items do contribute by proving some information towards the measurement of the ECT.

7.3.1.4 Misfit statistics

The misfit statistics allow one to explore the persons and items that are not behaving in the expected way. Table 22 presents the misfit order. The infit and outfit MNSQ values should be below 1.3 (Bond & Fox, 2007; Linacre & Wright, 2003; Pensavalle & Solinas, 2013), or they will be classified as problematic items. According to this criterion, the following items have outfit MNSQ values greater than 1.3 and are consequently cause for concern: item 17, item 12, item 16, item 18, and item 19. Item 37 has an outfit MNSQ values lower than 0.7 (Bond & Fox, 2007; Pensavalle & Solinas, 2013). Since there are no infit MNSQ values that are poor fitting, the items fit the model. The outfit MNSQ values, however, indicate that the items listed above are problematic and are outliers.

The point bi-serial measure correlation consists of a correlation between the total item score and the item score. This correlation indicates whether the item fits the abilities of the candidate. This is indicated by the strength of the correlation: The higher the correlation, the higher the item loading, which essentially indicates higher candidate measures for the latent variable (Linacre, 2011; Maree, 2004b, 2004c). Based on the correlation values observed, two items have problematic correlations, namely item 17 (.5) and item 16 (.5). The majority of the correlations are between low and moderate values, implying that there are possibly multiple constructs being measured by the test. Thus, the problematic and low correlations require further investigation to determine why they are not contributing more to the measure. The evidence of no negative correlations indicates that there are no reverse-scored items. This is to be expected, as the test does not contain any reverse-scored items. It should be noted that the primary reason for using this correlation is to establish that there is no incorrect coding or any negative values. These correlations will not be analysed due to their unusual distributions.

Table 22: Misfit Order Statistics for ECT version 1.2

Item	MNSQ Infit	MNSQ Outfit	Point Biserial Correlation (Real)	Point Biserial Correlation (Expected)
17	1.10	1.79	.05	.23
1	1.01	1.70	.15	.21
16	1.29	1.46	.05	.37
18	1.19	1.38	.11	.33
19	1.24	1.30	.11	.36
6	1.10	1.28	.21	.34
11	1.14	1.27	.19	.36
15	1.14	1.23	.20	.36
5	1.10	1.19	.26	.37
1	1.08	1.13	.28	.37
23	1.03	1.12	.31	.35
33	1.05	1.09	.29	.34
7	1.05	1.04	.31	.36
21	1.00	1.04	.32	.33
3	.99	1.02	.34	.33

8	1.02	1.00	.25	.27
14	1.02	1.01	.35	.37
26	.97	1.00	.38	.36
10	.99	.88	.17	.14
20	.99	.96	.37	.36
27	.98	.99	.37	.35
4	.98	.98	.37	.34
22	.98	.96	.39	.37
34	.96	.87	.27	.23
35	.96	.84	.31	.26
25	.94	.80	.32	.24
2	.92	.88	.43	.34
30	.92	.92	.45	.37
9	.91	.87	.48	.37
13	.91	.74	.31	.20
24	.91	.85	.45	.35
38	.89	.85	.49	.37
32	.88	.86	.49	.36
29	.86	.72	.48	.32
36	.86	.74	.47	.32
28	.83	.78	.53	.34
31	.83	.73	.51	.32
39	.83	.71	.51	.33
37	.81	.58	.52	.30

When considering the expected correlation according to the model, one can see from the column next to the point measure correlation that most of the correlations are similar to those expected. There are even cases where the expected was lower than the actual correlation. This indicates that the correlations are within the expected range for the model. There are, however, expected correlations that are much higher than the actual correlations and this causes some concern. This table also makes it easier to observe misfitting items (Linacre, 2011).

7.3.1.5 Dimensionality

The dimensionality of measurement instruments is essential to the exploration of the construct validity. Exploring dimensionality in the Rasch analyses is different from common factor analyses but is likened to principal components analyses (Linacre, 2012d). The interpretation is different, but the concept of exploring the constructs inherent in the measurement is shared. Since unidimensionality is one of the objectives of this study, it is necessary to explore the dimensionality of the ECT. In Rasch analyses, this takes the form of exploring the variance decomposition of the observations made, the standardised residuals contrast plots and standardised residual loading.

7.3.1.5.1 The variance decomposition of observations

In Table 23, the variance explained by the measures is 17 eigenvalues (31%), which indicates that the measure does not explain most of the variance in the ECT version 1.2. This could also suggest that there is not a wide spread of items and persons with different abilities, which implies that there is a similarity of difficulties and abilities or possible redundancy. These issues were observed in the item-person map. The raw unexplained variance is 39 (68%), which also indicates that the measure does not explain much of the variance in the ECT. This is, however, not concerning as the empirical data fit the model in terms of the predicted variance (Linacre, 2015, pp. 388-391). The model is the criterion against which the empirical data is compared. A small variance is, however, indicative of the quality of the test and indicates that the candidate's performance on the test does not provide for a wide variety of abilities. Thus, most candidates completing this test had similar abilities. There is also little unexplained variance.

When the eigenvalues are larger than 1.40, they suggest that there is an additional sub-dimension present. The unexplained variance in the first contrast is 3.08 eigenvalues (5%), in the second contrast is 2.1 eigenvalues (4%), in the third contrast is 1.58 eigenvalues (3%), and in the fourth contrast is 1.42 eigenvalues (3%). The presence of three dimensions indicates that the ECT is a multidimensional test.

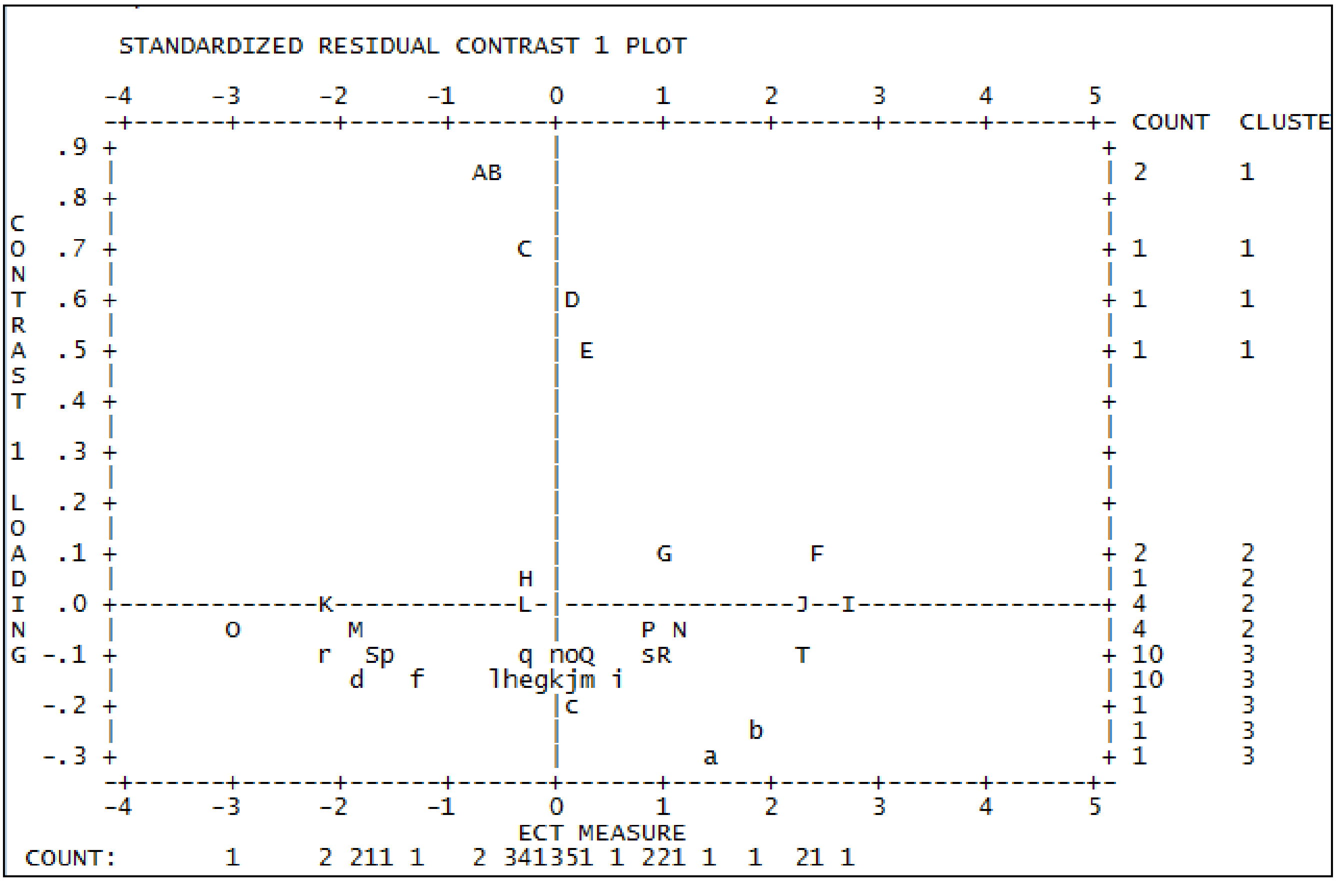
Table 23: Variance Decomposition of the Observations for the ECT version 1.2

	Empirical			**Modelled**
Variances	**Eigenvalues**	**Percentage**	**Percentage**	**Percentage**
Total raw variance in observations	56.6020	100%		100%
Raw variance explained by measures	17.6020	31.1%		31%
Raw variance explained by persons	5.495	9.7%		9.7%
Raw Variance explained by items	12.1066	21.4%		21.4%
Raw unexplained variance (total)	39.0000	68.9%	100%	68.9%
Unexplained variance in 1st contrast	3.0832	5.4%	7.9%	
Unexplained variance in 2nd contrast	2.1037	3.7%	5.4%	
Unexplained variance in 3rd contrast	1.5864	2.8%	4.1%	
Unexplained variance in 4th contrast	1.4222	2.5%	3.6%	

7.3.1.5.2 Standardised residual contrast plots

In Figure 14, the different dimensions are demonstrated with the different items belonging to each dimension. This is observed in the clusters in which these items appear in the plot. Figure 14 below shows the plot of the standardised residuals. The strength of the first contrast in the plot below is 3.08 eigenvalues, which is essentially 3 items. The standardised residual contrast plot assists one in exploring the local independence of items, which is one of the fundamentals of Rasch analyses (Linacre, 2012d). In this plot, the items that contribute to this contrast can be observed in the top left section of the plot. These three items are A, B, and C, as they are close together and within the same section of the plot. It can also be seen that items D and E are in the opposite section of the plot and these items are relatively further from the other items. This suggests that they are quite different from the other items in the ECT.

The difficulties of items are observed by viewing them on the *x*-axis from left to right. The size of the item loadings are indicated on the *y*-axis (Linacre, 2012d). The largest positive loading belongs to both items A and B. The largest negative loading belongs to item a, which can be found in the lower right corner. The items A and B, seem to contrast with the items a and b, in terms of their loadings and difficulties of items.



```
       STANDARDIZED RESIDUAL CONTRAST 1 PLOT

      -4     -3     -2     -1      0      1      2      3      4      5
      -+------+------+------+------+------+------+------+------+------+- COUNT  CLUSTE
   .9 +                            |                                   +
      |                       AB   |                                   | 2      1
   .8 +                            |                                   +
C     |                            |                                   |
O  .7 +                          C |                                   + 1      1
N     |                            |                                   |
T  .6 +                            |D                                  + 1      1
R     |                            |                                   |
A  .5 +                            | E                                 + 1      1
S     |                            |                                   |
T  .4 +                            |                                   +
      |                            |                                   |
1  .3 +                            |                                   +
      |                            |                                   |
L  .2 +                            |                                   +
O     |                            |                                   |
A  .1 +                            |      G         F                  + 2      2
D     |                          H |                                   | 1      2
I  .0 +-------------K------------L-|---------------J--I----------------+ 4      2
N     |       O       M            |     P N                           | 4      2
G -.1 +             r  Sp        q noQ   sR        T                   + 10     3
      |               d   f    lhegkjm i                               | 10     3
  -.2 +                            |c                                  + 1      3
      |                            |            b                      | 1      3
  -.3 +                            |         a                         + 1      3
      -+------+------+------+------+------+------+------+------+------+-
      -4     -3     -2     -1      0      1      2      3      4      5
                                 ECT MEASURE
 COUNT:       1     2 211 1   2 341351 1 221 1  1  21 1
```

Figure 14: Standardised Residual Contrast of the ECT version 1.2

7.3.1.5.3 Standardised residual loadings

The standardised residual contrast plot provides only limited information and thus more comprehensive information is required. In Table 24, the items and their respective loadings in the dimensions are displayed.

Table 24: Standardised Residual Loading for ECT version 1.2 (Sorted by Loading)

ECT Dimension	Item	Loading
1	29	.85
1	31	.85
1	28	.71
1	32	.58
1	30	.52
2	36	.08
2	38	.08
2	2	.03
2	37	.01
2	39	.01
2	13	.00
3	1	-.29
3	23	-.24
3	15	-.20
3	17	-.17
3	6	-.16
3	8	-.16
3	27	-.16
3	3	-.15
3	14	-.15
3	7	-.14
3	11	-.14

3	18	-.14
3	19	-.14
3	20	-.12
3	24	-.12
3	35	-.12
3	33	-.11
3	12	-.10
3	16	-.09
3	21	-.09
3	25	-.09
3	5	-.08
3	26	-.08
3	9	-.07
2	10	-.05
2	22	-.05
2	34	-.04
2	4	-.02

In Table 24, the different item loadings and three dimensions suggest that these clusters of items are measuring different constructs. There are only a few items which make up some of the dimensions and thus one may not necessarily consider them separate constructs. They could merely be elements of the construct being measured by the ECT; in other words, elements of verbal reasoning.

7.3.1.6 Characteristic curves

A Rasch analysis allows one to plot the items in a characteristic curve graphically. This curve indicates how the test and the items are performing compared to the Rasch model. Along with this, the information that the test gives is essential to explore. The graphical output provides a visual representation of the pattern the data form. The characteristic curves that will be explored are the test characteristic curve, the items characteristic curve, and the test information function curve.

7.3.1.6.1 Test characteristic curves

The test characteristic curve (TCC) for the ECT version 1.2 is shown below (Figure 15). This curve is interpreted by examining the shape of the line. This line indicates the expected score on the ECT in terms of the measurement of the latent variable. The graph shows an s-shaped line, which demonstrates that there is a fair fit between the items and the model. The steepness of the graph indicates that there is a relative range of difficulty, which implies that the items are relatively well spread out or distributed along the continuum (there are however some redundant items which affect the distribution). This steepness is also indicative of over-fit items in the test; thus some items are not yielding new information. The graph indicates that the poorer candidates score on the ECT, the less they will demonstrate of the latent variable and the lower their probability is of success. This is what one would expect, as higher scoring candidates should demonstrate more of the latent variable. This also indicates that the test is discriminating between the high and low performers in terms of item difficulty.

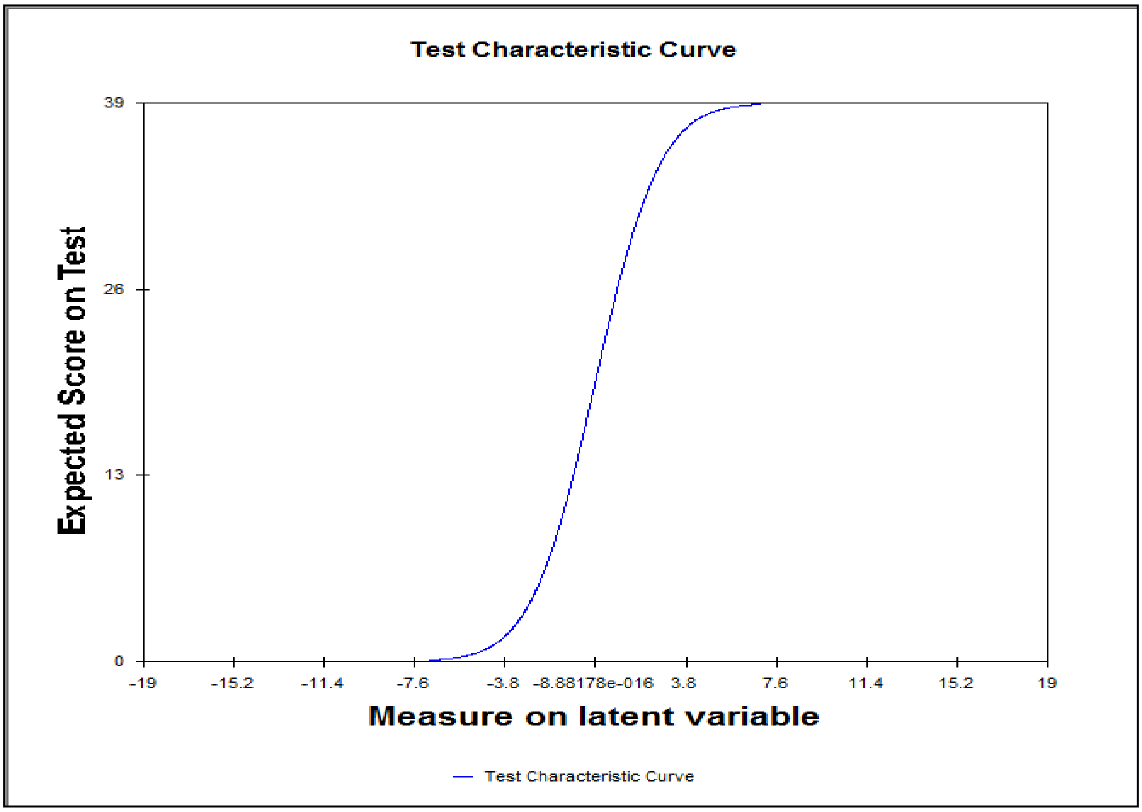


Figure 15: Test Characteristic Curve of the ECT version 1.2

7.3.1.6.2 Item characteristic curves

The item characteristic curve (ICC) indicates the probability of a person endorsing the specific item depending on where it is located according to its difficulty. The ICC curve of under-fitting items would be fairly flat, while the over-fitting items would have a steeper curve. The ICC can be produced for each item of the test (Linacre, 2012b, 2012c). This is, however, a tedious process, and with 39 items in the ECT version 1.2, it may become a redundant process and lose significance. For this reason, an ICC was produced for three items that are at varying difficulty levels. This allows one to observe the difference between easy, moderate, and difficult item patterns.

In Figure 16 below, the three lines shown are for the three items identified. The solid lines show the Rasch modelled ICC and the other skewed lines represented the actual empirical ICC. The model shows how the items should have looked if they performed ideally and if the responses were consistent. The points on the empirical ICC indicate the actual performance of the candidates on these items. In the first figure (Figure 16), the difference between the model and empirical ICC for the three items are not easily detected. The second ICC (Figure 17) containing only the empirical curves allow one to observe the difference more easily. This indicates the difference between the actual performance and the model ICC. This also shows the weaker candidates versus the more able candidates' performance on items.

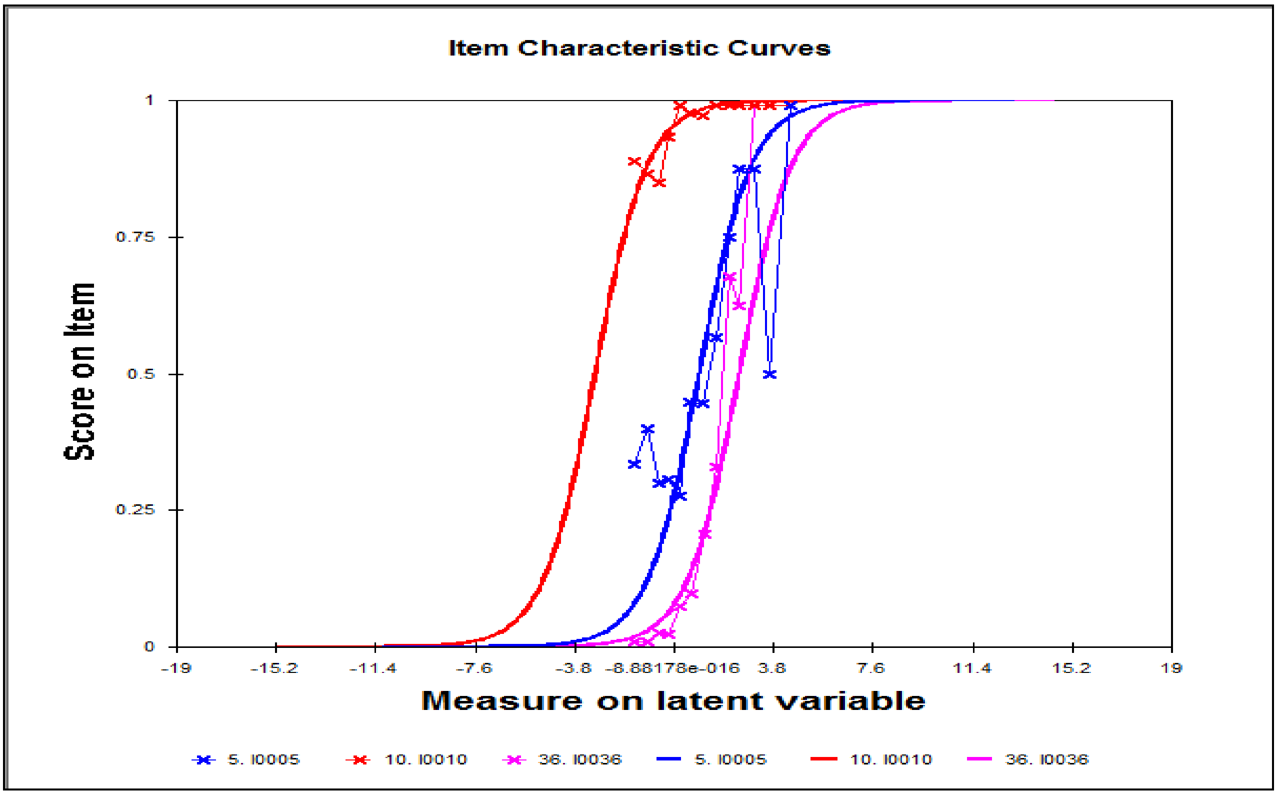


Figure 16: Item Characteristic Curves for Items 5, 10, and 36 (ECT version 1.2)

The easy item, item 10, is indicated by the red line and has the greatest contrast to the model ICC in terms of the performance on this item. Most candidates experience item 10 as very easy, and most were able to correctly answer this item, except for a few higher ability candidates. The moderate difficulty item, item 5, is indicated by the blue line and contrasts slightly with the model ICC. Item 5 was correctly answered by lower ability candidates, and this increased with ability, except for some incorrect responses by higher ability candidates. Item 36 is the difficult item and is indicated by the pink line. Item 36 mostly follows the curve of the model ICC, in that lower ability candidates will answer this item incorrectly and higher ability candidates will answer correctly.

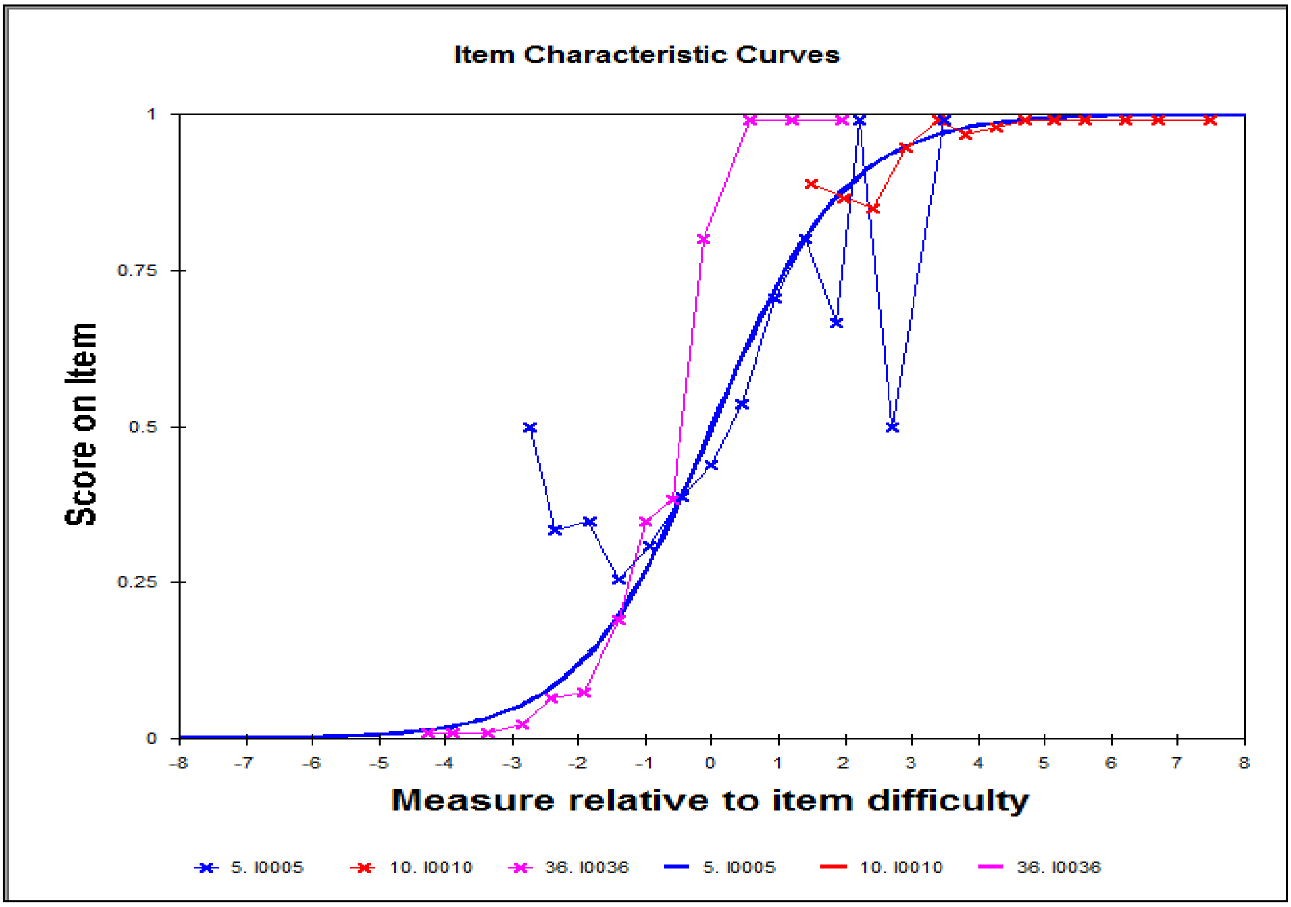


Figure 17: Item Characteristic Curves for Items 5, 10, and 36 (ECT version 1.2)

7.3.1.6.3 Test information function curve

The test information function indicates the point at which the most precise information of the test is found. This information is produced by the Fisher information function on the test items across the latent variable. The values and the shape of the graph are crucial when evaluating the test information function. The graph needs to have sufficient width, which indicates that there is effective measurement range in the test. The peak is important as it indicates critical cut-off points, specifically for criterion-related tests. The peak also represents the mode of the sample (Linacre, 2016; Yu, 2013).

In Figure 18 below, the graph shows a relatively pointy distribution, which centres on the mean. The points at which the most accurate information in the test will be found are between 7.6 and -7.6 logits. This indicates that there are almost equal amounts of easy and difficult items measuring the latent variable. The information yielded by the measure on the latent variable peaks at 7.5 logits. The width of this graph suggests that there is a relative measurement range in the test.

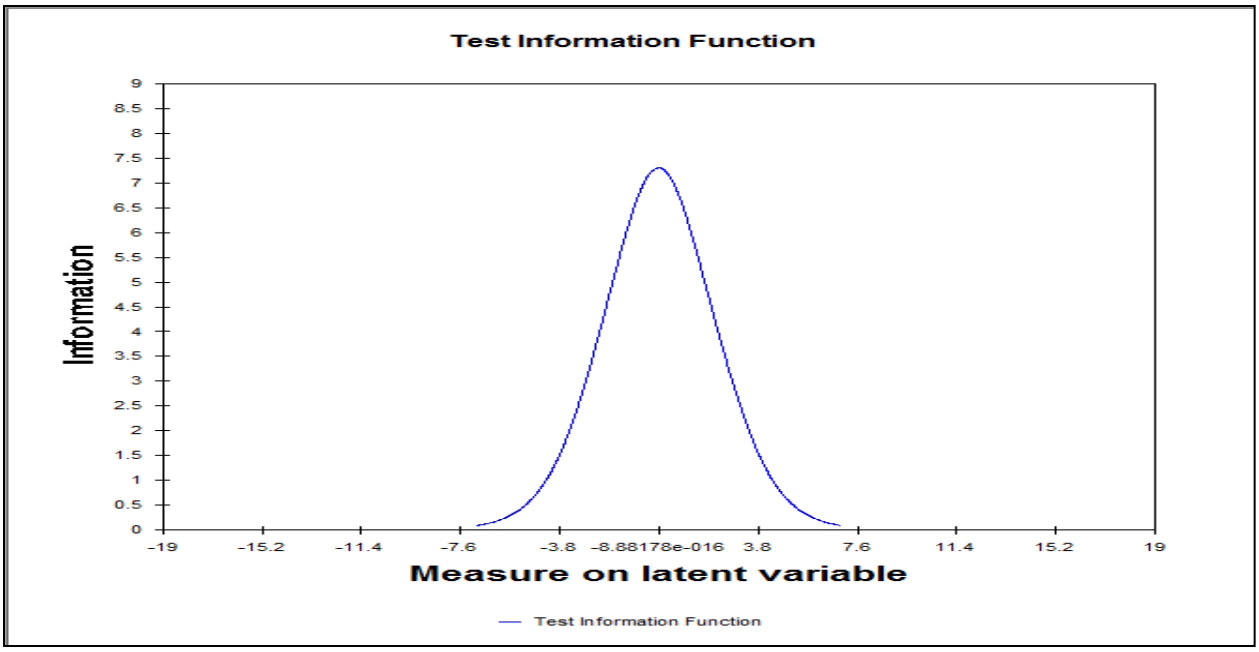


Figure 18: Test Information Function Curve for the ECT version 1.2

7.3.2 ECT Version 1.3 Results

The output will be presented in the following order: the fit statistics (the person statistics, the item statistics and test empirical randomness); the summary of category structure statistics; the person-item map; the measure order statistics (the bubble chart); the misfit order statistics; dimensionality (the variance decomposition of observations, the standardised residual contrast plots, the standardised residual loadings), and the characteristic curves (the test characteristic curve, the item characteristic curve and the test information function curve).

7.3.2.1 Fit statistics

7.3.2.1.1 The person statistics

Table 25: Average Person Statistics for ECT version 1.3

	Total Score	**Measure**	**MNSQ Infit**	**MNSQ Outfit**	**ZSTD Infit**	**ZSTD Outfit**
Mean	25.9	.69	1.00	1.02	.0	.0
Max	39	3.54	1.93	9.90	3.9	4.9
Min	8	-2.03	.55	.36	-2.6	-2.0

Tables 25 and 26 indicate the model fit information. The average person statistics are presented in Table 25. The mean is 25.9, which is the same as the mean in the initial descriptive statistics results (chapter 7, section 7.2.4). The average infit MNSQ value is 1.00 and the average outfit MNSQ value is 1.02, which is considered good and implies that on average, the person’s abilities fit the model. The average infit and outfit ZSTD is acceptable for both and thus is not statistically significant, which indicates that there is no misfit on average (Maree, 2004b, 2004c).

The maximum infit MNSQ value is 1.93 and the maximum outfit MNSQ value is 9.90. These values are both outside of the appropriate criteria and indicate that there are irregular patterns within the person's abilities. Since these values are relatively high, it suggests that quite a few persons do not fit the model. Further, the high outfit value implies that there are outliers that are negatively impacting the model. The minimum infit MNSQ value is 0.55 and the minimum outfit MNSQ value is 0.36. These values indicate that there are problems with persons in the model and this should prompt an investigation into the model to clarify which persons do not fit the model. The maximum (3.9 and 4.9) and minimum (-2.6 and -2.0) infit and outfit ZSTD values are very large and are therefore statistically significant. These large ZSTD values indicate that there are persons who are misfitting. The candidate reliability value is .77, which is not a very good reliability value as it indicates to some error in the observed score. The person separation value is 1.81, which is considered rather small and signifies that there is limited variation in the person's abilities. The standard error of the candidate mean is 0.03, which is very small, suggesting minimal error is observed (Linacre, 2016). The standard error of measurement value is 2.59 which can also be considered small and implies a small amount of error in the average measurement of a person (Linacre, 2011, 2012b; 2012c; Maree, 2004b, 2004c).

7.3.2.1.2 The item statistics

Table 26: Average Item Statistics for ECT version 1.3

	Measure	MNSQ Infit	MNSQ Outfit	ZSTD Infit	ZSTD Outfit
Mean	.00	.99	1.02	-.1	.2
Max	3.77	1.27	2.06	8.8	9.9
Min	-2.85	.78	.51	-4.7	-5.3

The average item statistics are indicted in Table 26. The infit MNSQ value is 0.99 and the outfit MNSQ value is 1.02, which indicates that on average, the items fit the model. The mean infit and outfit ZSTD values are acceptable and indicate that there are no items misfitting, as they are statistically insignificant. The maximum infit MNSQ value is 1.27 and the maximum outfit MNSQ value is 2.06, which suggests that there are items that are not behaving as expected in the model. The minimum infit MNSQ value is 0.78 and the minimum outfit MNSQ value is 0.51. The infit value is within range, while the outfit value is relatively low. This would indicate that there are outliers and possible noise within the items that should be investigated. The maximum (8.8 and 9.9) and minimum (-4.7 and -5.3) infit and outfit ZSTD are quite large and are statistically significant, indicating that there are items that are severely misfitting in the ECT. The item separation value is 15.71, which is considered a large value and denotes that there is good variation of item difficulties in the test. This also indicates that there should be a broad distribution observed for the items in the item map. The standard error is 0.24, which is small and suggests that there is little error observed (Linacre, 2011, 2012b; 2012c; Maree, 2004b, 2004c).

7.3.2.1.3 Test empirical randomness

Figure 19 below shows the test empirical randomness (Linacre, 2015). The MNSQ infit deviates only slightly towards the lower measure of the ECT. The MNSQ outfit, however, deviates significantly from the expected randomness, which occurs towards the mean measure of the ECT. This would suggest that the test is affected by severe outliers, such as external factors, which may have impacted the candidate’s performance on the test. This confirms the results found in the person and item statistics, in that there was a very large outfit MNSQ observed.

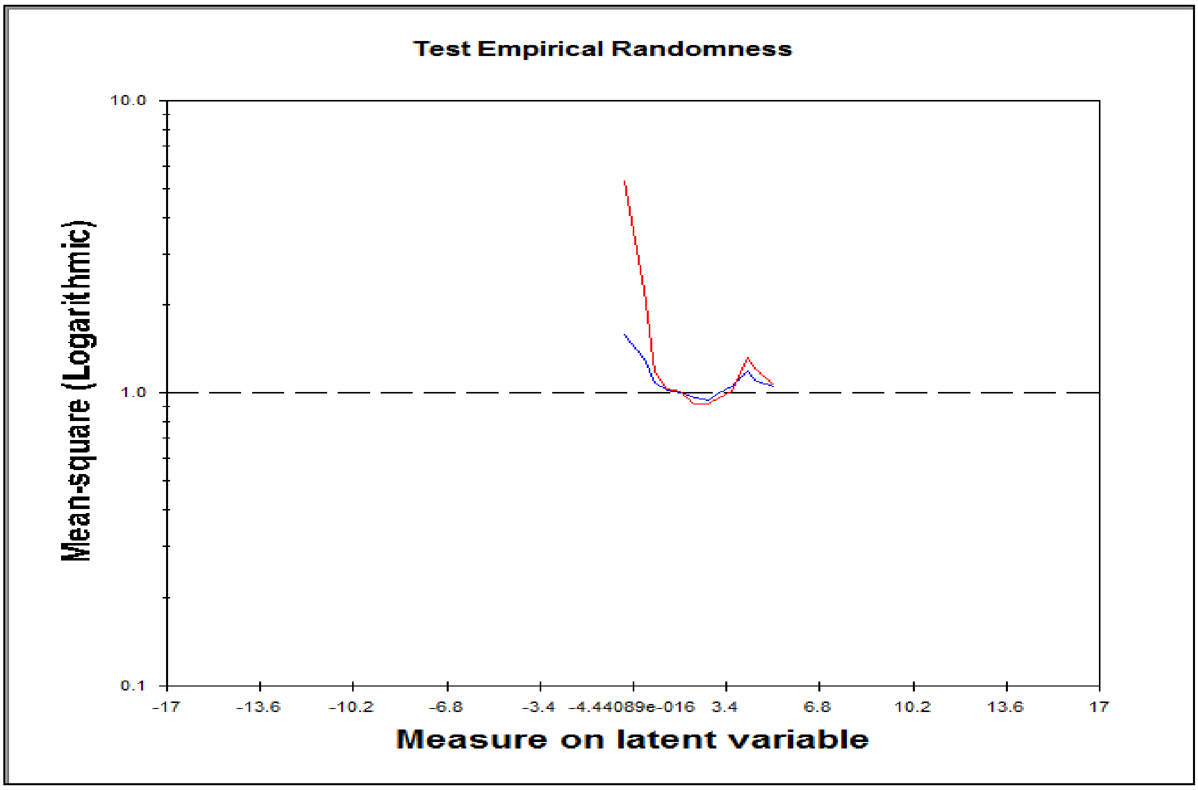


Figure 19: Test Empirical Randomness of the ECT version 1.3

7.3.2.1.4 Summary of category structure statistics

The summary of the category structure shown in Table 27 below, presents the values attributed to each of the categories. Category 0 comprised 38% (14163) of the total responses for the test, while category 1 comprised 62% (22839) of total responses.

Table 27: Summary of Category Structure Statistics for ECT version 1.3

Category label	**Observed Average**	**Expected Average**	**MNSQ Infit**	**MNSQ Outfit**
0	-.61	-.61	1.00	.97
1	1.50	1.50	1.00	1.09

It can be observed that more responses fell into category 1 than category 0. This means that most candidates were able to answer the questions in the test correctly. The average measures, increases vastly from - 0.61 to 1.50. Furthermore, the observation average values are equal to the expected values. The infit MNSQ values are 1.00 (category 0) and 1.00 (category 1), which are both acceptable infit values. The outfit MNSQ values for both categories are within range of 1. Category 1 = 1.09 and category 0 = 0.97, which are acceptable outfit values. This indicates that there is no noise or unexpected observations present in the responses for the test (Maree, 2004b).

```
        DICHOTOMOUS CURVES
P      -+--------------+--------------+--------------+--------------+-
R  1.0 +                                                             +
O      |                                                             |
B      |0                                                           1|
A      | 000000                                               111111 |
B   .8 +       00000                                     11111       +
I      |            0000                             1111            |
L      |                0000                     1111                |
I      |                    000               111                    |
T   .6 +                       000         111                       +
Y      |                          000   111                          |
    .5 +                             ***                             +
O      |                          111   000                          |
F   .4 +                       111         000                       +
       |                    111               000                    |
R      |                1111                     0000                |
E      |            1111                             0000            |
S   .2 +       11111                                     00000       +
P      | 111111                                               000000 |
O      |1                                                           0|
N      |                                                             |
S   .0 +                                                             +
E      -+--------------+--------------+--------------+--------------+-
       -2             -1              0              1              2
     Candidate [MINUS] ECT 1.3 MEASURE
```

Figure 20: Probability Curves of Observations in the two Categories of the ECT version 1.3

The probability curve of observation for the two response categories of the ECT can be seen in Figure 20. Category 0 ranges from just below 0.9 to 0.1, while category 1 ranges from 0.1 to 0.9. These probability curves indicate that low ability candidates (those below the mean ability) have fewer correct responses and more responses that are incorrect. The higher ability candidates (those above the mean), however, have fewer incorrect responses and more responses that are correct (Linacre, 2012c; Maree, 2004b).

7.3.2.2 Person-item map

```
  MEASURE Candidate - MAP - ECT 1.3
           <more>|<rare>
4                +
                 |
                 |  I0023
                 |
              .  |
                 |
                 |
              .  |T
3                +
                 |  I0039  I0040
            .##  |
                 |
            ###  |  I0025  I0042
                T|
            .##  |
                 |
2        .#####  +
        .######  |  I0027
                 |
      .########  |  I0001
     .######### S|S
                 |  I0026
     .#########  |  I0008  I0041
   ############  |  I0005
1   .##########  +
                 |
   .########### M|
  .############  |
   .###########  |  I0012  I0024  I0028
   .###########  |  I0033  I0036
  .############  |  I0011
     .#########  |  I0015
0      ########  +M I0038
                S|
          .####  |  I0002  I0004  I0007  I0034
        .######  |  I0010  I0013
           .###  |  I0003  I0006  I0014  I0032
          .####  |
             .#  |  I0018
             .#  |  I0019  I0035  I0037
  -1            .# T+
                .#  |
                    |  I0016
                 .  |  I0009
                 .  |S I0017
                    |
                    |
                 .  |  I0031
  -2             .  +
                    |  I0021
                    |  I0029
                    |  I0022
                    |  I0020
                    |
                    |
                    |  I0030
  -3                +
              <less>|<freq>
 EACH "#" IS 5: EACH "." IS 1 TO 4
```

Figure 21: Person-Item Map for the ECT version 1.3

In the person-item map (Figure 21), it can be seen that there is generally a good fit between the persons and the model, and the items are well spread along the continuum. This is evidenced by the distribution of the items and persons according to difficulty. There are, however, a few gaps in the distribution which would require items of a particular difficulty level. The dotted rectangles indicate that several items are measuring the same difficulty level and are thus redundant. These items could have addressed the gaps in the distribution.

It is also evident that in most cases, the person's ability fits alongside the item difficulty. Thus, the items cover the abilities of most of the candidates. There are, however, a few items that are too difficult for the persons and one item falls below their ability. For the items that are above the difficulty level of the person's ability location, the person has a less than 50% probability of correctly answering the item (Dunne, Long, Craig & Venter, 2012; Long, 2011). The majority of the candidates are between -1 and 2 standard deviations. This would suggest that the persons who completed the ECT range from low ability to average and above average ability. The majority of these candidates are within the average ability level.

The items of the test, however, cover a range of difficulties, in that there are very easy, easy, moderate, above average, and high difficulty items that are distributed along the continuum.

7.3.2.3 Measure order statistics

The measure order statistics are indicated in Table 28. When observing the values in the total score column, the most difficult items are items 23, 40, 39, 25, and 42, as only 53, 110, 118, 146, and 156 individuals respectively out of 881 individuals were able to answer these items correctly. The easiest items were items 30, 20, 22, and 29, as 847, 835, 830, and 824 individuals respectively out of 881 answered these items correctly. It is interesting to note that there is no item that no individuals were able to answer, and not a single item that all individuals succeeded in answering correctly. In the measure column, one can observe that there are quite a few items above the mean of 0 logits, thus indicating that these items require a higher ability level (Maree, 2004b, 2004c). There are, however, also several items below the mean that require lower ability level items.

Table 28: Misfit Order Statistics for the ECT version 1.3

Item	Total Score	Total Count	Measure
23	53	881	3.77
40	110	881	2.90
39	118	881	2.81
25	146	881	2.54
42	156	881	2.45
27	230	881	1.89
1	270	881	1.63
26	314	881	1.37
41	327	881	1.30

8	332	881	1.27
5	368	881	1.07
12	482	881	.46
24	482	881	.46
28	485	881	.45
36	497	881	.39
3	505	881	.34
11	529	881	.21
15	553	881	.08
38	565	881	.01
2	606	881	-.23
4	612	881	-.26
7	614	881	-.28
34	620	881	-.31
13	623	881	-.33
10	632	881	-.39
6	640	881	-.44
14	644	881	-.46
3	646	881	-.48
32	655	881	-.54
18	678	881	-.70
37	702	881	-.88
19	705	881	-.90
35	709	881	-.93
16	740	881	-1.20
9	763	881	-1.43
17	775	881	-1.56
31	802	881	-1.91
21	815	881	-2.12
29	824	881	-2.28
22	830	881	-2.41
20	835	881	-2.52

30	847	881	-2.85

7.3.2.3.1 The bubble chart

In Figure 22 below, the bubble chart for the ECT version 1.3 is shown (Linacre, 2015). When viewing the chart, the items on the top, such as items 39, 40, 42, and 23, are the most difficult items, and the items at the end, such as items 30, 20, 22, and 29, are the easiest. This corresponds to the previous table. The items that have either more (larger bubble) or less (smaller bubble) measurement errors are shown below (Bond & Fox, 2007; Linacre, 2012a; Pae et al., 2012).

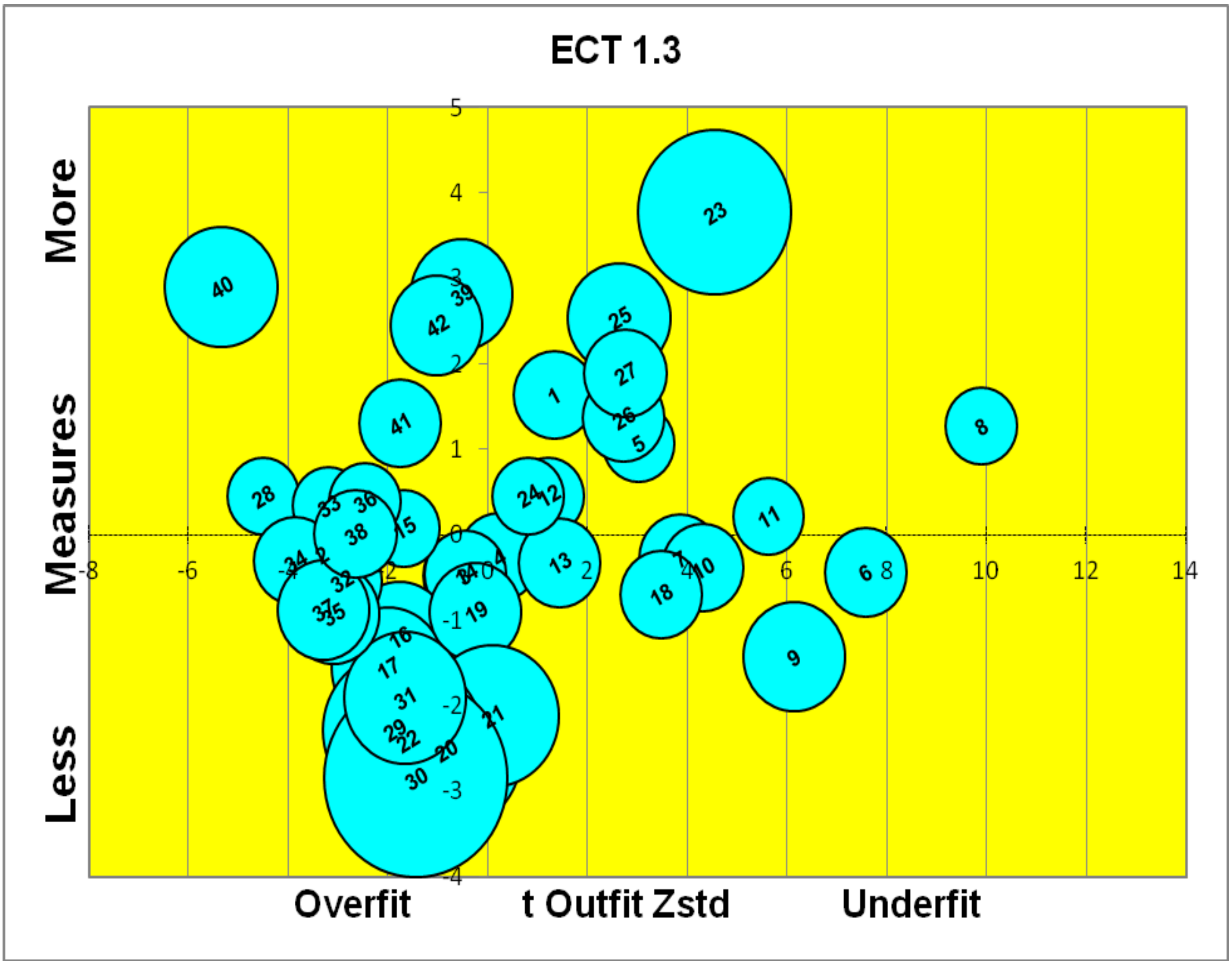


Figure 22: Bubble Chart for the ECT version 1.3

The easiest item (item 30) and the difficult items (item 39, 40, 42, and 23), which are much bigger bubbles compared to the other bubbles are seen in the extreme ends of the graph (Bond & Fox, 2007). The bubbles in the centre of the chart are relatively the same size and cluster close to each other. This would suggest that the items in the centre were more accurate, in terms of the measurement error, at measuring the candidate's ability in the ECT than the items situated on the extreme top and bottom. One would expect the tails to have less precise estimates, thus large errors. All the items are indicated in the bubbles, while some are not as easily legible when compared to others. This allows one to conclude that all the items do contribute by providing information about the test.

7.3.2.4 Misfit statistics

Table 29 presents the misfit order. According to the criterion (Bond & Fox, 2007; Linacre & Wright, 2003; Pensavalle & Solinas, 2013), the following items have outfit MNSQ greater than 1.3 and are consequently concerning: item 23, item 9, item 6, and item 8. Items 30 and 40 have outfit MNSQ values lower than 0.7. There are no infit MNSQ values that are poor fitting; no items fall below 0.7 and none fall above 1.3 (Bond & Fox, 2007; Pensavalle & Solinas, 2013), thus the items fit the model. The outfit MNSQ values, however, indicate that the items listed above are problematic and are outliers.

Table 29: Misfit Order Statistics for the ECT version 1.3

Item	MNSQ Infit	MNSQ Outfit	Point Biserial Correlation (Real)	Point Biserial Correlation (Expected)
23	1.07	2.06	.06	.21
9	1.13	1.83	.03	.26
6	1.14	1.51	.13	.33
8	1.27	1.46	.04	.37
10	1.16	1.26	.15	.33
18	1.12	1.26	.17	.31
25	1.03	1.26	.23	.30
11	1.19	1.23	.15	.36
7	1.11	1.22	.20	.34
27	1.03	1.17	.28	.34
26	1.06	1.12	.29	.36
5	1.07	1.11	.29	.37

13	1.04	1.08	.29	.34
1	.99	1.07	.36	.36
12	1.04	1.04	.33	.37
24	1.02	1.03	.35	.37
4	1.00	1.01	.34	.34
21	.97	1.01	.23	.21
19	1.00	.98	.30	.30
14	.99	.97	.34	.33
3	.95	.97	.37	.33
16	.97	.84	.33	.28
17	.96	.78	.32	.25
31	.96	.77	.29	.22
42	.95	.91	.38	.31
20	.94	.83	.25	.18
32	.94	.83	.41	.33
36	.94	.92	.44	.37
39	.93	.94	.36	.28
41	.94	.93	.43	.37
15	.93	.93	.43	.36
30	.93	.69	.25	.16
22	.92	.72	.29	.19
29	.91	.70	.31	.20
33	.91	.89	.46	.36
38	.91	.89	.45	.35
2	.89	.84	.46	.34
28	.89	.85	.49	.37
35	.88	.77	.44	.30
34	.87	.80	.48	.34
37	.87	.76	.45	.30
40	.78	.51	.54	.28

The point measure correlation indicates that there are three items that have problematic correlations, namely item 23, item 9, and item 8. The majority of the correlations are moderate values, although there are a few low correlations observed. This could suggest that there are multiple constructs being measured by the test. These correlations require further examination to determine why they are not adding more to the measure. There are no negative correlations because there are no reverse-scored items. Some of the actual correlations are higher and lower than the expected correlations, which is a cause of concern. Overall, the results indicate that the correlations are generally within expected range for the model (Linacre, 2011).

7.3.2.5 Dimensionality

The dimensionality analyses involves exploring the variance decomposition of the observations made, the standardised residuals contrast plots and standardised residual loadings.

7.3.2.5.1 The variance decomposition of observations

In Table 30, the variance explained by the measures is 23 eigenvalues (36%), which indicates that the measure explains some of the variance in the ECT version 1.3. This could be indicative of the fact that there is a limited range of items and persons with different difficulties and abilities, which implies that there is possible redundancy. This corresponds to the observations made on the item-map. The raw unexplained variance is 42 (65%), which also indicates that the measure explains very little of the variance in the ECT. It is also noted that the empirical data fits the model in terms of the predicted variance (Linacre, 2015, pp. 388-391). The small variance indicates that most candidates completing this test had similar abilities. Since the eigenvalues are larger value 1.40, they suggest that there are additional sub-dimensions present. The unexplained variance in the first contrast is 2.90 eigenvalues (5%), in the second contrast is 2.13 eigenvalues (3%), in the third contrast is 1.68 eigenvalues (3%), in the fourth contrast is 1.52 eigenvalues (2%), and in the fifth contrast is 1.35 eigenvalues (2%). The presence of four sub-dimensions (based on the cut-off of 1.40 eigenvalues in a dimension) indicates that the ECT is a multidimensional test.

Table 30: Variance Decomposition of the Observations for the ECT version 1.3

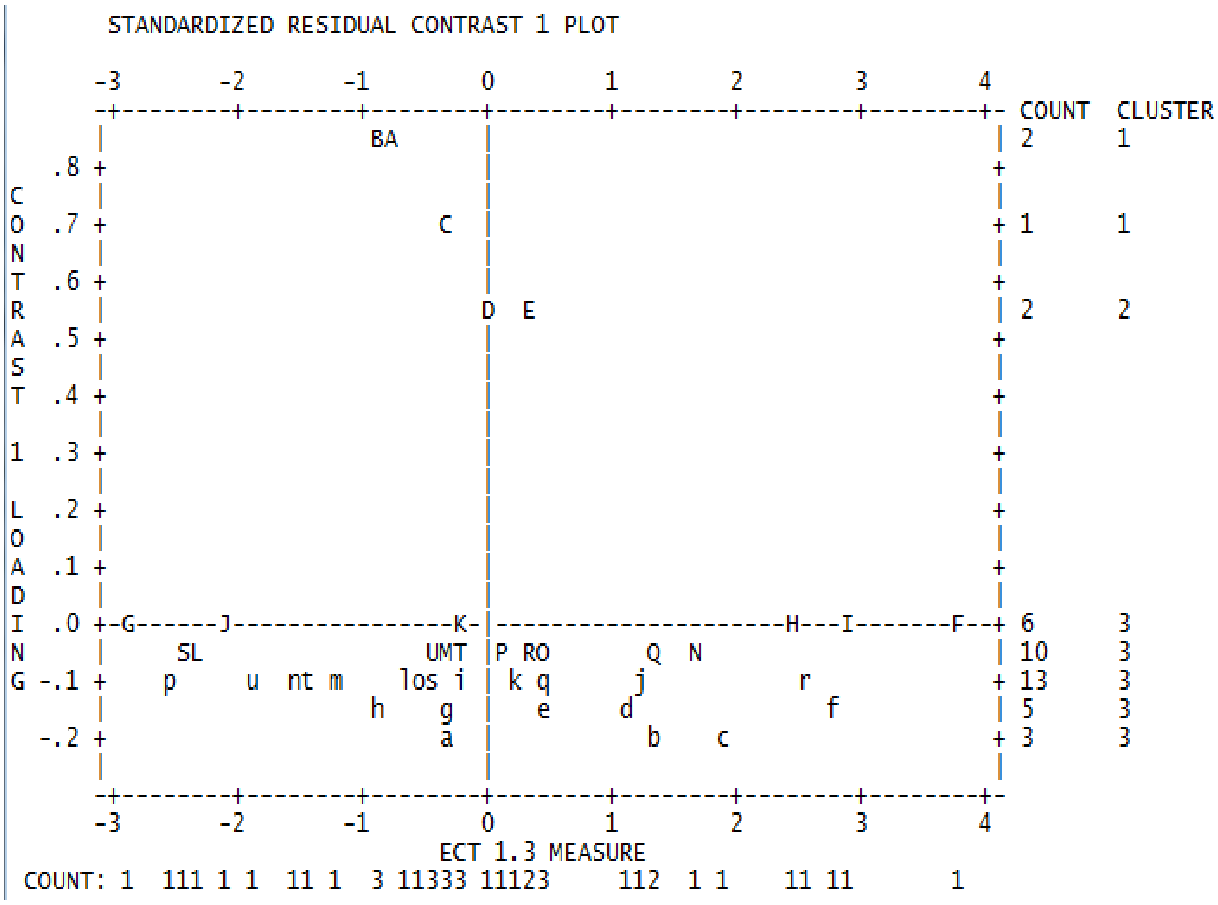
	Empirical			**Modelled**
	Eigenvalues	**Percentage**	**Percentage**	**Percentage**
Total raw variance in observations	65.1646	100%		100%
Raw variance explained by measures	23.1646	35.5%		35.5%
Raw variance explained by persons	6.4432	9.9%		9.9%
Raw Variance explained by items	16.7213	25.7%		25.6%
Raw unexplained variance (total)	42.0000	64.5%	100%	64.5%
Unexplained variance in 1st contrast	2.9010	4.5%	6.9%	
Unexplained variance in 2nd contrast	2.1370	3.3%	5.1%	
Unexplained variance in 3rd contrast	1.6820	2.6%	4.0%	
Unexplained variance in 4th contrast	1.5223	2.3%	3.6%	
Unexplained variance in 5th contrast	1.3501	2.1%	3.2%	

7.3.2.5.2 Standardized residual contrast plots

In Figure 23, the plot of the standardised residuals is shown. The strength of the first contrast in the plot below is 2.89 eigenvalues, which is essentially 3 items. In this plot, the items that contribute to this contrast can be observed in the top left section of the plot. These three items are B, A, and C, as they are close together and within the same section of the plot. It can also be seen that item D is on the mean line and item E is on right-hand side of the plot.

These items (A, B, C, D, and E) are much further from the other items. This suggests that these items differ from the other items in the ECT.

The largest positive loading belongs to both items B and A. The largest negative loadings belong to items a, b, and c which can be seen in the lower left and right corner. Items, B, A, and C seem to contrast with items, a, b, and c in terms of their loadings and difficulty levels (Linacre, 2012d).



```
        STANDARDIZED RESIDUAL CONTRAST 1 PLOT

       -3        -2        -1         0         1         2         3         4
       -+---------+---------+---------+---------+---------+---------+---------+- COUNT  CLUSTER
       |                     BA       |                                        | 2      1
    .8 +                              |                                        +
C      |                              |                                        |
O   .7 +                           C  |                                        + 1      1
N      |                              |                                        |
T   .6 +                              |                                        +
R      |                              D  E                                     | 2      2
A   .5 +                              |                                        +
S      |                              |                                        |
T   .4 +                              |                                        +
       |                              |                                        |
1   .3 +                              |                                        +
       |                              |                                        |
L   .2 +                              |                                        +
O      |                              |                                        |
A   .1 +                              |                                        +
D      |                              |                                        |
I   .0 +-G-------J------------------K-|-----------------------H----I--------F--+ 6      3
N      |     SL                  UMT  |P RO        Q   N                       | 10     3
G  -.1 +     p     u  nt  m    los  i | k q       j            r               + 13     3
       |                     h     g  |   e      d                f            | 5      3
   -.2 +                           a  |            b     c                     + 3      3
       |                              |                                        |
       -+---------+---------+---------+---------+---------+---------+---------+-
       -3        -2        -1         0         1         2         3         4
                                  ECT 1.3 MEASURE
  COUNT: 1  111  1 1  11  1  3 11333  11123      112  1  1    11 11         1
```

Figure 23: Standardised Residual Contrast of the ECT version 1.3

7.3.2.5.3 Standardized residual loadings

In Table 31, the items and their respective loadings in the dimensions are displayed. There are three dimensions identified with a few items indicated for the first two of the three dimensions. The item loadings across the three dimensions range from high to very low indicating that there are possibly different dimensions being measured. These dimensions could however be considered as dimensions of the construct of verbal reasoning.

Table 31: Standardised Residual Loading for ECT version 1.3 (Sorted by Loading)

ECT Dimension	Item	Loading
1	37	.85
1	35	.84
1	34	.70
2	38	.55
2	36	.54
3	23	.00
3	30	.00
3	13	-.20
3	26	-.20
3	27	-.19
3	5	-.17
3	12	-.16
3	39	-.14
3	10	-.13
3	19	-.13
3	7	-.12
3	8	-.12
3	11	-.12
3	18	-.12
3	16	-.11
3	17	-.11
3	32	-.11
3	20	-.10
3	24	-.10

3	25	-.10
3	14	-.09
3	9	-.08
3	31	-.08
3	3	-.06
3	4	-.06
3	22	-.06
3	33	-.06
3	41	-.06
3	15	-.05
3	28	-.05
3	1	-.04
3	6	-.04
3	29	-.03
3	2	-.01
3	21	-.01
3	40	-.01
3	42	-.01

7.3.2.6 Characteristic curves

7.3.2.6.1 Test characteristic curves

The TCC for the ECT version 1.3 is shown below (Figure 24). The graph shows an s-shaped line, indicating a reasonable fit between the items of the test and the model. The steepness of the graph indicates that there is a moderate range of difficulty for the items of the test, due to some redundant items in the distribution. These redundant items are shown in the person-item map. Furthermore, the item difficulty is able to discriminate between the high and low performers within the test.

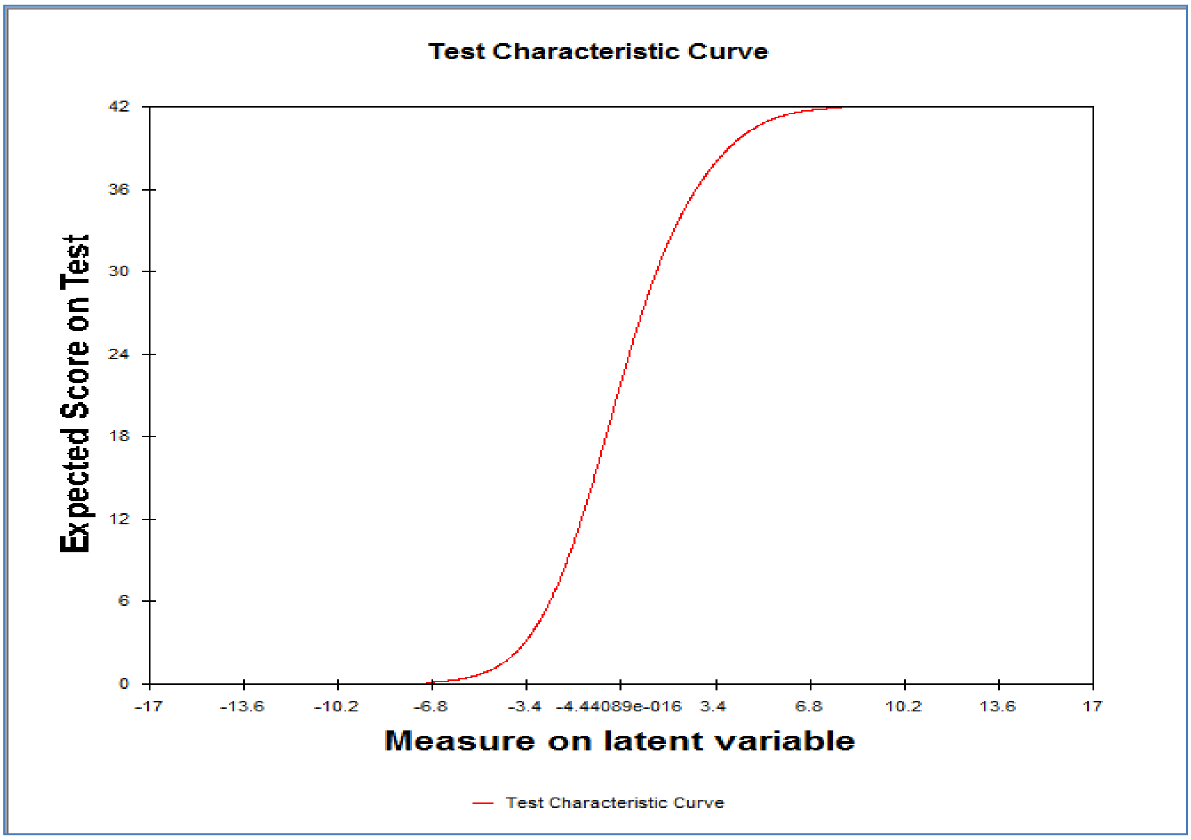


Figure 24: Test Characteristic Curve for the ECT version 1.3

7.3.2.6.2 Item characteristic curves

In Figure 25 below, the model and empirical ICC for the three items are indicated, while the second ICC (Figure 26) only indicates the empirical curves. For all three items indicated in the ICC, the lowest score was 0, which indicates that the very poor performing candidates were not able to answer any of these three items.

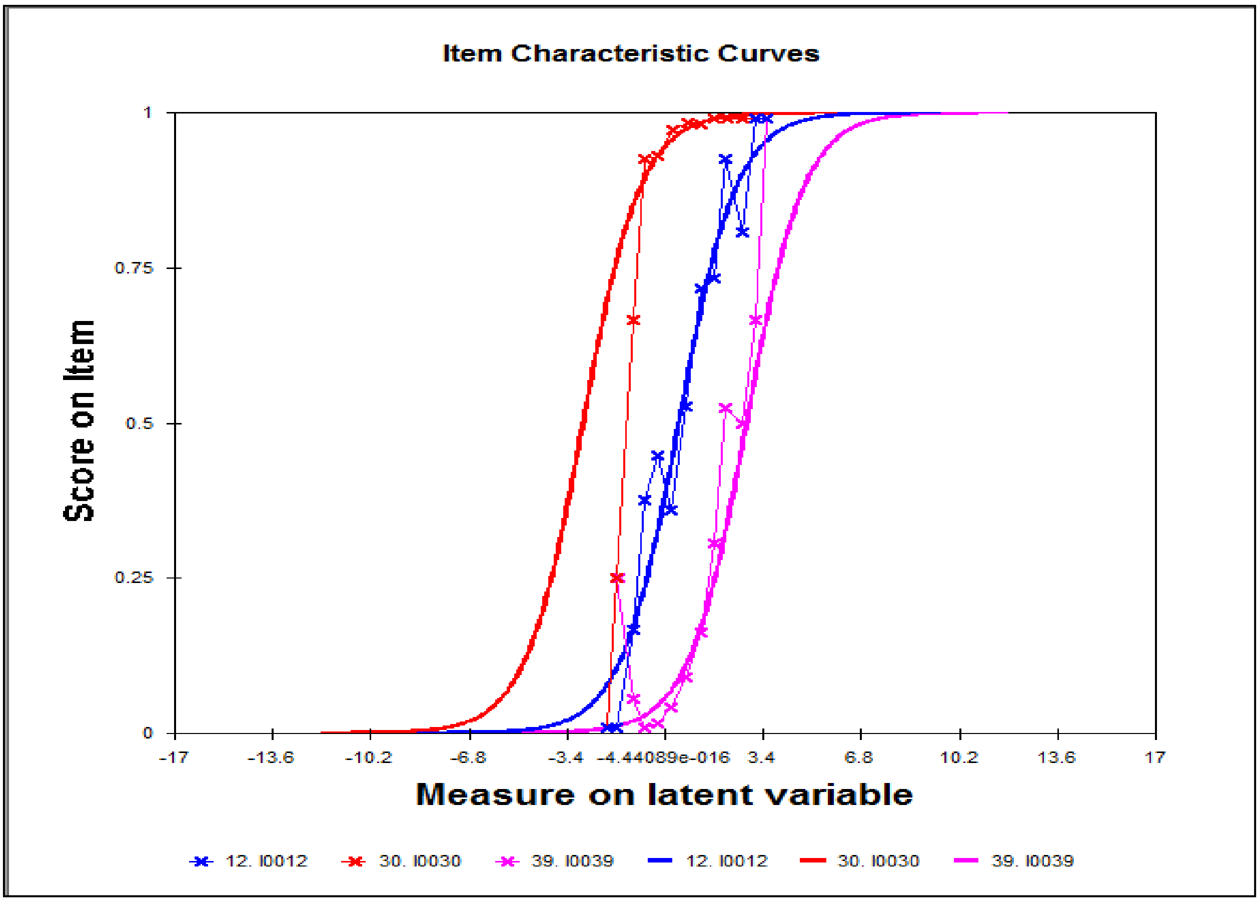


Figure 25: Item Characteristic Curve for the ECT version 1.3

The easier item, item 30, is indicated by the red line and contrasts only slightly with the model ICC. Generally, this item follows the ICC model curve pattern. This suggests that as the score increases on the item, the probability of poor candidates answering correctly decreases. Item 30 was experienced as difficult for the very poor candidates but as the measure increases, the score on this item increases so that most can answer this item correctly. There are mostly high scores on item 30, with the exception of only a few low scores.

The moderate difficulty item, item 12, is indicated by the blue line and contrasts slightly with the model ICC. Item 12 was incorrectly answered by very low ability candidates and this increased with ability as higher ability candidates are able to correctly answer item

12. There are average scores on item 12, with some high scores on this item. Item 39 is a very difficult item and is indicated by the pink line. Item 39 contrasts with the curve of the model ICC, in that most candidates answered this item incorrectly and only some high ability candidates are able to answer this item correctly. The performance on item 39 initially has a spike but overall it appears to increase with ability. This could be indicative of the fact that item 39 was experienced as difficult for most candidates.

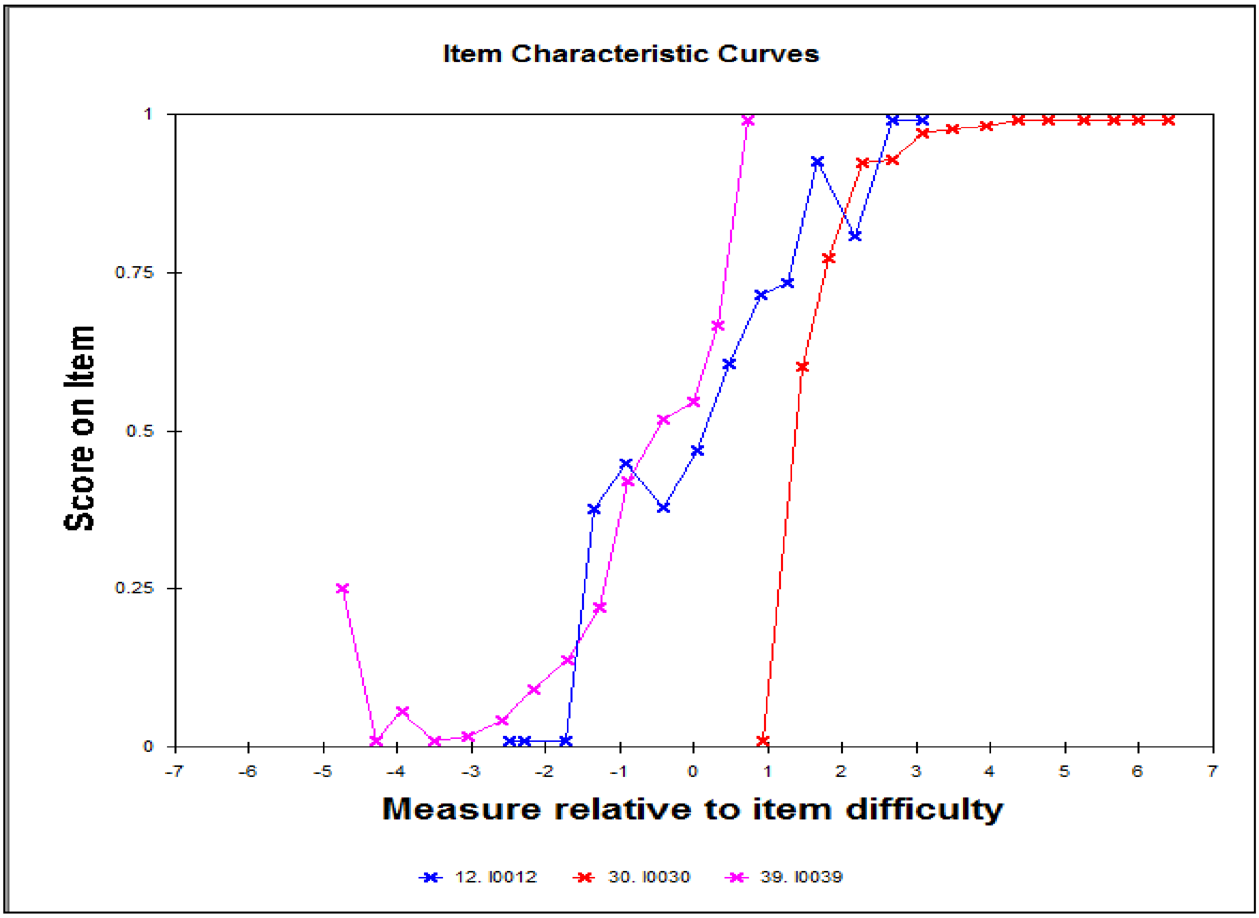


Figure 26: Item Characteristic Curve for the ECT version 1.3

7.3.2.6.3 Test information function curve

In Figure 27 below, the graph shows a pointy distribution, which is centred on the mean. The width of the graph indicates that there is a relative measurement range in the test and the information peaks at just over 7 logits. The points at which the most accurate information in the test will be found are between -6.8 and 7 logits. This indicates that candidates have a fair probability of success of endorsing easy items, while also having a fair probability of endorsing difficult items.

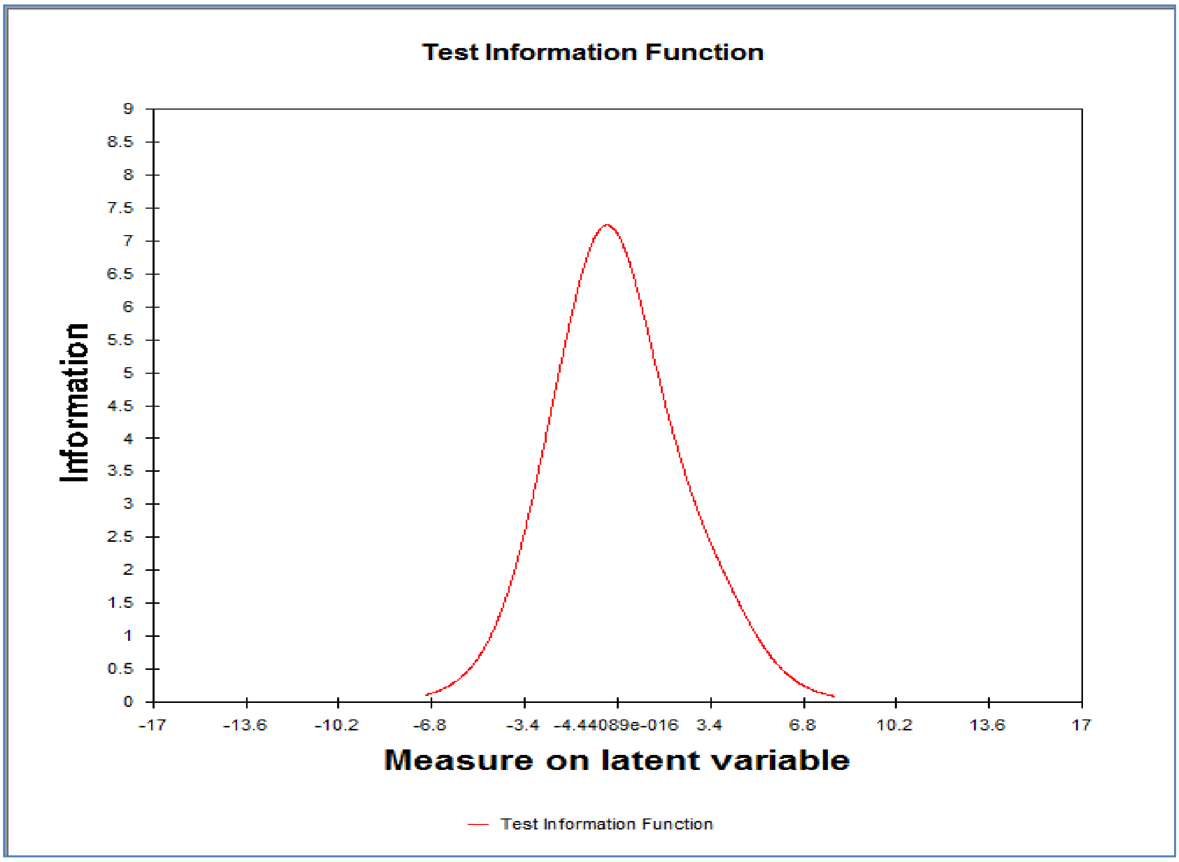


Figure 27: Test Information Function Graph for the ECT version 1.3

7.4 Confirmatory Factor Analysis

The confirmatory factor analysis was performed by conducting structural equation modelling (SEM) in AMOS, as explained in Chapter 6. The outputs of the SEM results were as follows: the graphical input of the ECT model, the assessment of normality, the regression weights, the squared multiple correlation, the standardised regression weights, the model fit statistic (chi-square statistic), the baseline comparison, the root square error of approximation (RSMEA), the root mean square residual (RMR) and the Akaike’s information criterion (AIC). These results will be presented only for the ECT Version 1.3.

7.4.1 The ECT Version 1.3 Results

In Figure 28 below, the graphical input of the ECT version 1.3 is displayed in terms of the hypothesized model the researcher wanted to confirm. This model was based on the exploratory factor analysis (EFA) performed on the ECT version 1.2 (Arendse & Maree, 2017). The labelled factors that were based on the loadings of the EFA (PAF) for the ECT version 1.2 were as follows: factor 1: Vocabulary, factor 2: Reasoning and factor 3: Deduction. These three factors are identified within the graphical input of the model and were also discussed in Chapter 4, section 4.7. The model was therefore created to confirm whether these factors, which were identified in the EFA of the first test version (ECT version 1.2), were confirmed as factors in the model of the second test version (ECT version 1.3).

The model used for the ECT was based on the EFA results of the first version of the test (ECT version 1.2). The factors of the model were specified before the CFA was conducted to make the model identifiable. This pre-specification of the model was based on the EFA results for ECT version 1.2 and applied to the second test version of the ECT (ECT version 1.3). The factors in the model are not specified as having a causal relationship to the latent variable of verbal reasoning. These factors (deduction, plurals, vocabulary, reasoning and education) are instead elements which make up the construct of verbal reasoning.

There are two additional factors that were added to complete the model of the ECT version 1.3. The first additional factor was labelled plurals, as these items were not present in the test version of the ECT 1.2. These items were therefore only in the test version of the ECT 1.3 and needed to be added to the CFA analyses. The second additional factor added was labelled education, based on the fact that none of these items loaded on the other factors (reasoning, deduction and vocabulary) in the ECT version 1.2. The education factor in this model and the items of which it is comprised could perhaps be influenced by educational factors, schooling, and level of knowledge, since they did not load on any of the other factors. The formulation of this factor would therefore also be evaluated in the CFA.

The latent variable, Verbal Reasoning, is theorized to be the underlying construct that is being measured by the different factors within the ECT. Verbal reasoning, as previously argued in Chapter 4, specifically 4.7, could be the underlying construct of the observed factors (vocabulary, reasoning (this refers to a general reasoning ability), deduction, plurals and education) of the ECT. The confirmation of the model will assist in the argument for unidimensionality of the ECT as well as providing evidence in support of the construct validity of the ECT.

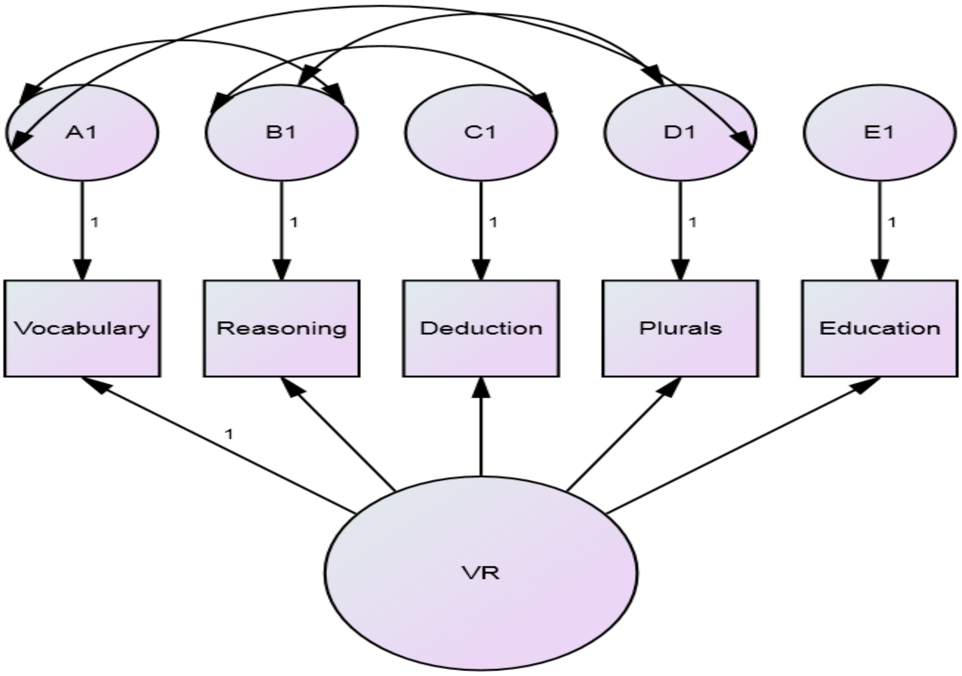


Figure 28: Graphical Input of the ECT version 1.3 Model

This graphical input can be described in the following way: Verbal reasoning (depicted as VR in the circular format) is the latent variable that the ECT is hypothesised to be measuring. Within this verbal reasoning construct are factors such as vocabulary, reasoning, deduction, plurals, and education. There are relationships among the following factors: vocabulary and plurals, vocabulary and reasoning, reasoning and deduction, and reasoning and plurals. There is, however, no relationship hypothesised by the four factors (vocabulary, reasoning, deduction and plurals) and the fifth factor, education. This model is unidimensional and complies with the three-indicator rule (Kline, 2011) previously discussed in Chapter 6, section 6.7.2.2 and 6.7.2.2.2.

Since this model was accepted by SPSS AMOS and was able to run, it suggests that the relationship that was hypothesised was also consistent with the sample data and was possibly the correct model of the ECT 1.3. It is, however, not sufficient to merely graphically explore the model as the model statistics are important and inform one whether or not to accept the model.

Table 32: Assessment of Normality for the ECT version 1.3

Variable	**Min**	**Max**	**Skew**	**Critical Ratio (C.R).**	**Kurtosis**	**C.R.**
Education	3.000	21.000	-.436	-5.284	.286	1.731
Plurals	.000	5.000	-1.170	-14.177	1.161	7.031
Deduction	.000	4.000	-.887	-10.748	1.262	7.648
Reasoning	.000	7.000	.717	8.687	.137	.832
Vocabulary	.000	5.000	-1.081	-13.096	-.164	-.994
Multivariate					5.619	9.967

In Table 32, the normality of the structure of the ECT version 1.3 is assessed. Based on this table, the minimum and maximum values for the education factor differ significantly more than those of the other factors. The skewness values for the different factors indicate that education, plurals, deduction, and vocabulary are negatively skewed, while the reasoning factor is positively skewed. Since skewness affects the means of the test, it is less important for SEM (Byrne, 2010). As 7 is the value used to indicate a deviation from normality for kurtosis (Kline, 2005, as cited in Byrne, 2010), the kurtosis for the factors indicates that all factors are essentially normally distributed, as none of them violate the cut-off of 7. The importance of assessing the kurtosis lies in the fact that it is impacted by the analysis of variances and covariances (Byrne, 2010). The *z* statistic of the critical ratio is used for interpretation of the multivariate kurtosis value (Byrne, 2010). The *z* statistic value of 9.967

suggests that there is a deviation from normality, as this value is greater than the cut-off value of 5 (Byrne, 2010).

Table 33: Regression Weights for the ECT version 1.3

Factors			Estimate	SE	C.R.
Vocabulary	<---	VR	1.000		
Reasoning	<---	VR	1.213	.112	10.843***
Deduction	<---	VR	.447	.045	9.961***
Plurals	<---	VR	.645	.064	10.059***
Education	<---	VR	2.697	.280	9.618***

*p <.05, **p < .01, ***p < .001

The regression weights are indicated in Table 33. These weights are also referred to as factor loadings. The highest factor loadings on verbal reasoning were education, reasoning, and vocabulary. The standard errors for the different factor loadings are all relatively small (except vocabulary, which is not displayed); indicating minimal error was observed in the estimation of these factor loadings (Byrne, 2010). The critical ratio for the factor loadings indicates that all the factors (except vocabulary) are statistically different from zero, as their *z* statistics all exceed the cut-off of 1.96 (Byrne, 2010). It should also be noted that all the regression weights are significant ($p > 0.001$).

Table 34: Squared Multiple Correlations of ECT version 1.3

Factors	Estimate
Education	.641
Plurals	.269
Deduction	.289
Reasoning	.409
Vocabulary	.232

The correlation estimate values for the five manifest factors of the ECT 1.3 are indicated in Table 34 above. The highest correlations are for education (.641) and reasoning (.409), while the lowest correlations observed are for plurals (.269) and vocabulary (.232). This correlation can also be interpreted as the extent to which the factors explain the variance in the model, similar to how r^2 is interpreted in the regression for the model variance (Terblanche, 2015). This would suggest that the education factor accounts for 64% and the reasoning factor accounts for 41% of the variance in the model.

Table 35: Standardised Regression Weights for the ECT version 1.3

Factors			Estimate
Vocabulary	<---	VR	.482
Reasoning	<---	VR	.640
Deduction	<---	VR	.538
Plurals	<---	VR	.518
Education	<---	VR	.801

Table 35 above indicates the standardised regression weights of the model, which simply put, indicates the influence of the different factors (observed variables) on the latent variable of verbal reasoning. The unstandardised regression weights were statistically significant. The largest impact on verbal reasoning is observed by the education (.801) and reasoning (.640), as they explain 80% and 64% of the variance respectively. The smallest effect on verbal reasoning is observed by the vocabulary (.482), which explains 48% of the variance.

Table 36: Model Fit Statistic for the ECT version 1.3

Model Statistic	Value
Chi-square	0.638
Degrees of Freedom	1
Probability level (*p*)	0.424

According to the model fit statistics shown above (Table 36), the degree of freedom is 1, indicating a positive and very restrictive amount. This suggests that the model is over-identified (Morgan, 2015) and this is ultimately what one seeks to provide evidence in support of the validity of the ECT. The chi-square value of 0.638 is not statistically significant (as *p* > 0.001). This is a positive result as it implies that the observed variance and covariance matrix is consistent with the model variance and covariance. This also means that the null hypothesis is accepted.

Table 37: Baseline Comparison for the ECT version 1.3

Model	NFI Delta1	RFI rho1	IFI Delta2	TLI rho2	CFI
Default model	.999	.993	1.000	1.004	1.000
Saturated model	1.000		1.000		1.000
Independence model	.000	.000	.000	.000	.000

In Table 37 above, the baseline comparison is made by exploring the different model fit statistics. The CFI (comparative fit index) is 1.000, which is considered excellent or a perfect model fit (Schumacher & Lomax, 2010). The NFI (normed fit index) is 0.999, which is considered very good (Schumacher & Lomax, 2010). The IFI (incremental fit index) is

1.000, which is also good and indicates good fit. The TLI (Tucker-Lewis index) is 1.004, which is also considered a good fit. The RFI (relative fit index) is 0.993, which is considered a good fit. In essence, all these model indeces indicated that the model is a good fit.

Table 38: The Root Square Error of Approximation for the ECT version 1.3

Model	**RMSEA**	**LO 90**	**HI 90**	**PCLOSE**
Default model	.000	.000	.082	.764
Independence model	.311	.293	.328	.000

The RMSEA (root square error of approximation) is shown above (Table 38). The RMSEA value of 0.000 indicates good model fit. This suggests that the model fits the population. In terms of the 90% confidence intervals (LO 90 and HI 90) (Morgan, 2015), the RMSEA is situated between 0.000 and 0.082. This would still be considered a reasonable model fit as the HI 90 value is slightly above the 0.082 cut-off (Hair et al., 2009).

Table 39: The Root Mean Square Residual for the ECT version 1.3

Model	**RMR**	**GFI**	**AGFI**	**PGFI**
Default model	.008	1.000	.996	.067
Saturated model	.000	1.000		
Independence model	.896	.656	.484	.437

In Table 39, the root mean square residual (RMR) is displayed. The RMR value ranges from 0 to 1, and the closer the value is to 0, the better the fit to the model (Byrne, 2010). The RMR value of .008 implies that the model explains the correlations to within an average error of .008. This indicates that there is a very small error observed between the sample and hypothesised model fit. Furthermore, the RMR value of .008 indicates a good fit

(< 0. 05) between the sample covariance matrix and the model covariance matrix. The goodness of fit (GFI) and adjusted goodness of fit (AGFI) values range from 0 to 1, with a value closer to 1 being a better fit. These statistics compare the model to no model existing and consider parsimony as part of the comparison (Byrne, 2010). The GFI value of 1.000 and AGFI value of .996, therefore, indicates a good fit. The PGFI considers parsimony in its evaluation of the model as well as the parameters estimated, thus the model fit and parsimony of the model are signified by the PGFI value. The PGFI value of .067 can be considered an acceptable model fit.

Table 40: The Akaike's Information Criterion for the ECT version 1.3

Model	**AIC**	**BCC**	**BIC**	**CAIC**
Default model	28.638	28.831	95.573	109.573
Saturated model	30.000	30.206	101.716	116.716
Independence model	868.439	868.507	892.344	897.344

In Table 40, the Akaike's information criterion (AIC) and Bozdogan's consistent version of Akaike's information criterion (CAIC) are indicted. The BCC and BIC are not interpreted. The AIC and CAIC essentially evaluate the model, the parsimony in the model, and the estimated parameters. The AIC and CAIC are interpreted in the same way and are compared to the saturated and independence model. The AIC and CAIC should, therefore, have smaller values than these models for it to be considered a good model fit (Byrne, 2010). The AIC value of 28.638 and CAIC value of 109.573 are both smaller than the saturated and independence models, thus they indicate good model fit.

7.5 Multi-Trait Multi-Method Analyses

To assist in the interpretation of the MTMM analysis results, the results will be reported as follows: correlations of the ECT and the different psychometric tests (AAT, DAT, and SATs), the reliability diagonals within the psychometric tests, and the mono-trait mono-method triangles. These results are presented for the two test versions separately.

7.5.1 The ECT Version 1.2 Results

In Table 41 below, the correlations of the different AATs with the ECT are observed. The bolded diagonal includes the reliability values. The highest correlations are identified by the red rectangles marked in the table. The highest correlations for the ECT were observed between the following tests: ECT 1.2 and Verbal Reasoning, ECT and Vocabulary, and ECT and Reading Comprehension.

Table 41: Psychometric Test Comparisons of the AAT tests for the ECT version 1.2

Psychometric Tests	**ECT 1.2**	**Non-Verbal Reasoning AAT 2**	**Verbal Reasoning AAT 2**	**Vocabulary AAT 3**	**Reading Comprehension AAT 4**	**Numeric Comprehension AAT 5**
ECT 1.2	**.79**	.392**	.640**	.729**	.711**	.427**
Non-Verbal Reasoning	.392**	**.87**	.659**	.300**	.478**	.639**
Verbal Reasoning	.640**	.659**	**.76**	.616**	.729**	.694**
Vocabulary	.729**	.300**	.616**	**.85**	.775**	.395**
Reading Comprehension	.711**	.478**	.729**	.775**	**.81**	.489**
Numeric Comprehension	.427**	.639**	.694**	.395**	.489**	**.94**

*p <.05, **p < .01, ***p < .001

The highest correlations among the AAT were observed for the following tests: Non-Verbal Reasoning and Verbal reasoning, Vocabulary and Verbal Reasoning, Reading

Comprehension and Verbal Reasoning, Numeric Comprehension and Non-Verbal reasoning, Vocabulary and Reading Comprehension, and Verbal Reasoning and Numeric Comprehension. It should be noted that all the correlations for the Verbal Reasoning test are considered large correlations.

The reliability diagonal observed in Table 40 above indicates very strong reliabilities for the tests, ranging from over .70 to .90.

Table 42: Psychometric Test Comparisons of the DAT and SAT tests for the ECT version 1.2

Tests	**ECT 1.2**	**Verbal Reasoning DAT 2**	**Non-Verbal Reasoning DAT 3**	**Mechanical Insight DAT 9**	**Long-Term Memory DAT 10**	**Calculations SAT 2**	**Comparisons SAT 4**	**Pattern Completion SAT 5**	**Figure Series SAT 6**	**Spatial 2D SAT 7**	**Spatial 3D SAT 8**	**Short-Term Memory**
ECT 1.2	**.79**	.528**	.306**	.382**	.535**	.362**	.295**	.252**	.349**	.299**	.297**	.295**
DAT 2	.528**	**.55**	.584**	.480**	.624**	.520**	.307**	.426**	.594**	.443**	.479**	.459**
DAT 3	.306**	.584**	**.71**	.454**	.514**	.525**	.344**	.481**	.719**	.561**	.627**	.412**
DAT 9	.382**	.480**	.454**	**.54**	.513**	.473**	.251**	.364**	.475**	.561**	.558**	.383**
DAT 10	.535**	.624**	.514**	.513**	**.77**	.559**	.423**	.444**	.525**	.483**	.454**	.460**
SAT 2	.362**	.520**	.525**	.473**	.559**	**.921**	.423**	.425**	.571**	.496**	.492**	.368**
SAT 4	.295**	.307**	.344**	.251**	.423**	.423**	**.762**	.331**	.404**	.356**	.320**	.248**
SAT 5	.252**	.426**	.481**	.364**	.444**	.425**	.331**	**.834**	.490**	.433**	.409**	.379**
SAT 6	.349**	.594**	.719**	.475**	.525**	.571**	.404**	.490**	**.852**	.584**	.637**	.424**
SAT 7	.299**	.443**	.561**	.561**	.483**	.496**	.356**	.433**	.584**	**.918**	.666**	.454**
SAT 8	.297**	.479**	.627**	.558**	.454**	.492**	.320**	.409**	.637**	.666**	**.838**	.400**
SAT 10	.295**	.459**	.412**	.383**	.460**	.368**	.248**	.379**	.424**	.454**	.400**	**.762**

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

In Table 42 above, the highest correlations are indicated by the red rectangles. The diagonal of bolded values shows that the reliability values are relatively strong. These values range from .50 to .90. The majority of the reliability values are, however, relatively high, and it would indicate that there is a strong construct present.

The strongest (largest) correlations for the ECT and the DATs and SATs are between the ECT and Verbal Reasoning and the ECT and Long–Term Memory. The strongest correlations among the DAT and SAT are as follows: Verbal Reasoning and Long-Term Memory, Non-Verbal Reasoning and Figure Series, Non-Verbal reasoning and Spatial 3D, Spatial 3D and Figure Series, and Spatial 2D and Spatial 3D.

These psychometric tests (table 42) are not theorised to be related to the constructs of the ECT 1.2. Based on the correlations observed for the ECT 1.2, the following was observed. Firstly, the largest correlations observed for the ECT among these tests were for the Mechanical Insight, Calculations, and Figure Series tests, and range from .349 – .382. Secondly, the smallest correlations observed for the ECT amoung these tests were for Pattern Completion, Spatial 3D, and Spatial 2D, and range from .252 – .299. Thirdly, the largest correlations observed for the Figure Series tests were for the Non-Verbal, Spatial 3D, Verbal Reasoning, and Spatial 2D tests. Fourthly, the strongest correlations among most of the above tests were with the Figure Series and Spatial 2D tests. Fifthly, most of these tests are predominately non-verbal in nature and the highest correlations among them are with the non-verbal tests, with the exception of the Long-Term Memory and Verbal Reasoning tests. As observed in the correlations, one expects larger relationships between these non-verbal tests as they are theorised to be similar.

Since the traditional MTMM analysis could not be completed, as discussed in the Chapter 6 the modified multi-trait mono-method analysis was conducted in the form of mono-trait mono-method triangles (Figure 29). Since the AAT tests were not conducted at the same time or with the same people as those who completed the DAT and the SAT, the results cannot be compared to each other.

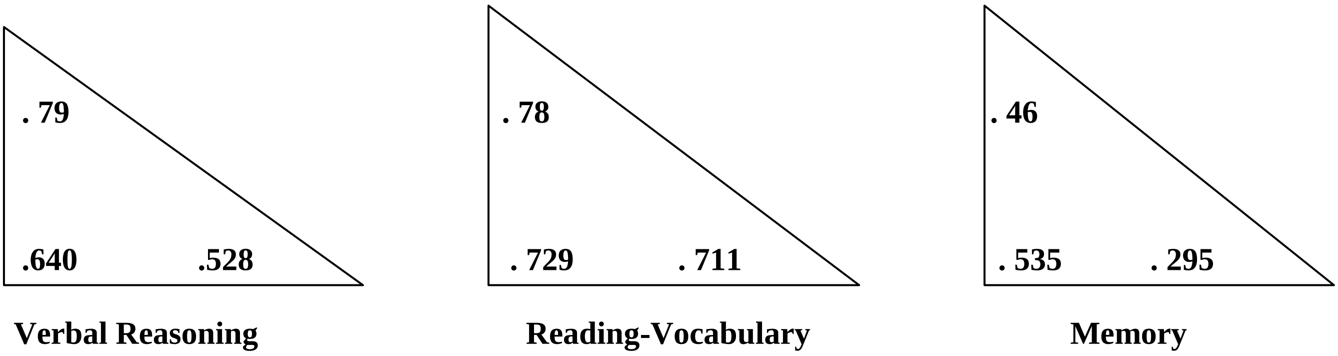


Figure 29: Mono-Trait Mono-Method Triangles for the ECT version 1.2

Since the ECT was, however, the common variable tested with these tests, the correlation of the ECT with these constructs will be explored. The verbal reasoning triangle of the mono-trait mono-method triangle (Figure 29) includes the correlation of the ECT and Verbal Reasoning tests (AAT 2 and DAT 2). The correlation of the AAT 2 and DAT 2 could not be compared because the sample was different, so the reliability of the ECT was placed in the triangle instead. All the correlations are considered large and indicate that there is a strong relationship between these constructs.

The reading-vocabulary reasoning triangle of the mono-trait mono-method triangle (Figure 29) includes the correlation of the ECT and Verbal Reasoning tests (AAT 3 and AAT 4). The correlation of the AAT 3 and AAT 4 is also shown (.78). All the correlations are considered large and indicate that there is a strong relationship between these constructs. The correlations observed in the reading-vocabulary reasoning triangle are the highest correlations observed for the ECT when compared with the correlations of the other triangles.

The memory-reasoning triangle of the mono-trait mono-method triangle (Figure 29) includes the correlation of the ECT and Verbal Reasoning tests (DAT 10 and SAT 10). The correlation of the DAT 10 and SAT 10 is also shown. The correlation between the long-term memory and the ECT is large, while there is a small correlation between short-term memory and the ECT. The correlation between the short-term and long-term memory is moderate (.46). These different sizes of correlations indicate a definite relationship between long-term memory and the ECT, while short-term memory and the ECT have a very small relationship. This is, however, expected since long-term and short-term memory has a moderate relationship.

7.5.2 The ECT Version 1.3 Results

In Table 43 below, the correlations of the different AATs with the ECT are observed. The highest correlations for the ECT version 1.3 were observed between the following tests: Verbal Reasoning, Vocabulary, and Reading Comprehension.

Table 43: Psychometric Test Comparisons of the AAT tests for the ECT version 1.3

Psychometric Tests	**ECT 1.3**	**Non-Verbal Reasoning AAT 2**	**Verbal Reasoning AAT 2**	**Vocabulary AAT 3**	**Reading Comprehension AAT 4**	**Numeric Comprehension AAT 5**
ECT 1.3	**.79**	.269**	.410**	.633**	.637**	.309**
Non-Verbal Reasoning	.269**	**.87**	.595**	.320**	.311**	.557**
Verbal Reasoning	.410**	.595**	**.76**	.536**	.553**	.535**
Vocabulary	.633**	.320**	.536**	**.85**	.703**	.362**
Reading Comprehension	.637**	.311**	.553**	.703**	**.81**	.339**
Numeric Comprehension	.309**	.557**	.535**	.362**	.339**	**.94**

*p <.05, **p < .01, ***p < .001

The highest correlations among the AAT were observed for the following tests: Non-Verbal Reasoning and Verbal Reasoning, Vocabulary and Verbal Reasoning, Reading Comprehension and Verbal Reasoning, Numeric Comprehension and Non-Verbal Reasoning, Vocabulary and Reading Comprehension, and Verbal Reasoning and Numeric Comprehension. All the correlations observed for the Verbal Reasoning test are large, suggesting a strong link to most of the tests.

Table 44: Psychometric Test Comparisons of DAT and SAT tests for the ECT version 1.3

Tests	**ECT 1.3**	**Verbal Reasoning DAT 2**	**Non-Verbal Reasoning DAT 3**	**Mechanical Insight DAT 9**	**Long-Term Memory DAT 10**	**Calculations SAT 2**	**Comparisons SAT 4**	**Pattern Completion SAT 5**	**Figure Series SAT 6**	**Spatial 2D SAT 7**	**Spatial 3D SAT 8**	**Short-Term Memory SAT 10**
ECT 1.3	**.79**	.604**	.386**	.405**	.580**	.455**	.293**	.303**	.386**	.344**	.354**	.418**
DAT 2	.604**	**.55**	.528**	.497**	.604**	.577**	.419**	.474**	.530**	.445**	.466**	.458**
DAT 3	.386**	.528**	**.71**	.511**	.438**	.501**	.392**	.495**	.670**	.592**	.598**	.426**
DAT 9	.405**	.497**	.511**	**.54**	.495**	.507**	.376**	.461**	.508**	.542**	.565**	.389**
DAT 10	.580**	.604**	.438**	.495**	**.77**	.543**	.441**	.417**	.469**	.409**	.405**	.529**
SAT 2	.455**	.577**	.501**	.507**	.543**	**.921**	.486**	.528**	.514**	.484**	.438**	.381**
SAT 4	.293**	.419**	.392**	.376**	.441**	.486**	**.762**	.421**	.421**	.370**	.332**	.362**
SAT 5	.303**	.474**	.495**	.461**	.417**	.528**	.421**	**.834**	.508**	.469**	.461**	.408**
SAT 6	.386**	.530**	.670**	.508**	.469**	.514**	.421**	.508**	**.852**	.576**	.561**	.463**
SAT 7	.344**	.445**	.592**	.542**	.409**	.484**	.370**	.469**	.576**	**.918**	.679**	.393**
SAT 8	.354**	.466**	.598**	.565**	.405**	.438**	.332**	.461**	.561**	.679**	**.838**	.392**
SAT 10	.418**	.458**	.426**	.389**	.529**	.381**	.362**	.408**	.463**	.393**	.392**	**.762**

*p <.05, **p < .01, ***p < .001

In Table 44 above, the largest correlations for the ECT 1.3 and the DATs and SATs; are between the ECT and Verbal Reasoning, and the ECT and Long–Term memory. There are also moderate correlations between the ECT and calculations and short-term memory. The strongest correlations among the DAT and SAT are as follows: Verbal Reasoning and Long-Term Memory, Verbal Reasoning and Calculations, Non-Verbal Reasoning and Figure Series, Non-Verbal Reasoning and Spatial 2D, Non-Verbal Reasoning and Spatial 3D, Spatial 2D and Figure Series, Spatial 3D and Figure Series, and Spatial 2D and Spatial 3D.

There were large correlations observed for the ECT among the Calculations, Mechanical Insight, Non-Verbal, and Figure Series tests, which ranged from .386 – .455. The smallest correlations observed for the ECT were amoung the Comparison, Pattern Completion, and Spatial 2D tests, and ranged from .293 – .344. The largest correlations observed for the Figure Series tests were with the Non-Verbal, Spatial 2D, Spatial 3D, and Verbal Reasoning tests. The Spatial 2D test had the largest correlations with these tests. Interestingly, the Long-Term Memory and Verbal Reasoning tests had large correlations with the non-verbal tests.

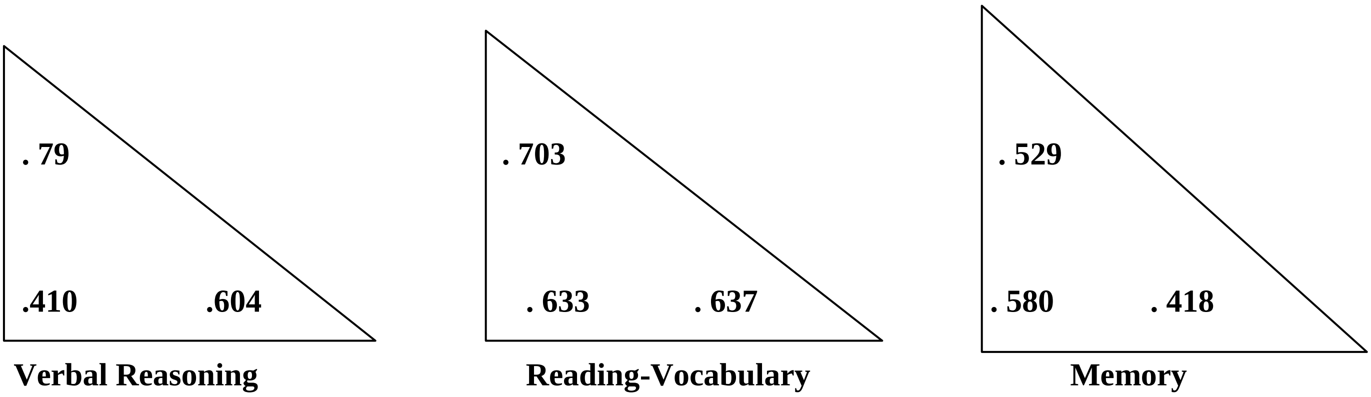


Figure 30: Mono-Trait Mono-Method Triangles for the ECT version 1.3

The verbal reasoning triangle (figure 30) has correlations that are moderate and large, indicating that there is a definite relationship between these constructs, yet not to the same extent as in the two subtests. The reading-vocabulary reasoning triangle (Figure 30) contains correlations that are considered large and suggests that there is a strong relationship between these constructs. The correlations for the reading-vocabulary reasoning triangle are the highest for the ECT when compared to the correlations in the other triangles.

The memory-reasoning triangle (Figure 30) shows that the correlation between the short-term memory and the ECT is large, while the correlation between short-term memory and the ECT is a moderate one. The correlation between the short-term and long-term memory is also moderate (.529). The sizes of these correlations indicate a definite relationship between long-term memory and the ECT, while short-term memory and the ECT have a moderate relationship. There is thus a strong relation of the ECT to memory.

7.5.3 Comparison of Psychometric Tests for ECT 1.2 and ECT 1.3

Table 45: Correlations for the Verbal Reasoning, Vocabulary, and Reading Comprehension Tests for the ECT version 1.2

Psychometric Tests	**ECT 1.2**	**Verbal Reasoning AAT 2**	**Vocabulary AAT 3**	**Reading Comprehension AAT 4**
ECT 1.2	.79	.640**	.729**	.711**
Non-Verbal Reasoning	.392**	.659**	.300**	.478**
Verbal Reasoning	.640**	.76	.616**	.729**
Vocabulary	.729**	.616**	.85	.775**
Reading Comprehension	.711**	.729**	.775**	.81
Numeric Comprehension	.427**	.694**	.395**	.489**

*p <.05, **p < .01, ***p < .001

Table 46: Correlations for the Verbal Reasoning, Vocabulary, and Reading Comprehension Tests for the ECT version 1.3

Psychometric Tests	ECT 1.3	Verbal Reasoning AAT 2	Vocabulary AAT 3	Reading Comprehension AAT 4
ECT 1.3	.79	.410**	.633**	.637**
Non-Verbal Reasoning	.269**	.595**	.320**	.311**
Verbal Reasoning	.410**	.76	.536**	.553**
Vocabulary	.633**	.536**	.85	.703**
Reading Comprehension	.637**	.553**	.703**	.81
Numeric Comprehension	.309**	.535**	.362**	.339**

*p <.05, **p < .01, ***p < .001

In the two tables above (Table 45 and Table 46), the three psychometric tests identified were Verbal Reasoning, Vocabulary, and Reading Comprehension. These tests were selected as they are theorised to have the greatest relation to the constructs of the ECT. Based on the correlations observed for the two test versions, the following observations can be made. Firstly, the correlations among these three psychometric tests (Verbal Reasoning, Vocabulary, and Reading Comprehension) became smaller from the ECT version 1.2 to the ECT version 1.3. This would suggest that the relationship still exists but for some reason was not as large as in the ECT version 1.3. This could be due to the persons completing the tests or the changes across the test versions (such as the time limit imposed on the ECT 1.2).

Secondly, the strongest correlations with the ECT across both versions were observed for the following tests: Vocabulary, Reading Comprehension, and Verbal Reasoning. Thirdly, the correlation between the ECT and Verbal Reasoning became smaller for the ECT version 1.3. Fourthly, the Vocabulary and Reading Comprehension tests have the strongest relationship with Verbal Reasoning. This was observed for both test versions. Fifthly, the strongest relationship observed for the Verbal Reasoning test was consistently with Reading

Comprehension across both test versions, which corresponds to literature (Lakin, 2012). Sixthly, the Vocabulary test had the strongest correlations with the Reading Comprehension and Verbal Reasoning tests. These correlations, therefore, indicate that the strongest relationship consistently emerged between the ECT and Verbal Reasoning, Vocabulary, and Reading Comprehension. This provides evidence of the theorised relationship between the constructs and confirms that the ECT has consistently tapped into these constructs, indicating the existence of a definite reasoning factor.

Table 47: Correlations for Verbal Psychometric Tests for the ECT version 1.2 and 1.3

Tests	ECT 1.2	Verbal Reasoning DAT 2	Long-Term Memory DAT 10	Short-Term Memory SAT 10	ECT 1.3	Verbal Reasoning DAT 2	Long-Term Memory DAT 10	Short-Term Memory SAT 10
ECT 1.2	**.79**	.528**	.535**	.295**	**.79**	.604**	.580**	.418**
DAT 2	.528**	**.55**	.624**	.459**	.604**	**.55**	.604**	.458**
DAT 3	.306**	.584**	.514**	.412**	.386**	.528**	.438**	.426**
DAT 9	.382**	.480**	.513**	.383**	.405**	.497**	.495**	.389**
DAT 10	.535**	.624**	**.77**	.460**	.580**	.604**	**.77**	.529**
SAT 2	.362**	.520**	.559**	.368**	.455**	.577**	.543**	.381**
SAT 4	.295**	.307**	.423**	.248**	.293**	.419**	.441**	.362**
SAT 5	.252**	.426**	.444**	.379**	.303**	.474**	.417**	.408**
SAT 6	.349**	.594**	.525**	.424**	.386**	.530**	.469**	.463**
SAT 7	.299**	.443**	.483**	.454**	.344**	.445**	.409**	.393**
SAT 8	.297**	.479**	.454**	.400**	.354**	.466**	.405**	.392**
SAT 10	.295**	.459**	.460**	**.762**	.418**	.458**	.529**	**.762**

*p <.05, **p < .01, ***p < .001

In Table 47 above, the different psychometric tests that were used during the piloting of the ECT 1.2 and ECT 1.3 are indicated. These psychometric tests are also theorised to be related to the constructs of the ECT 1.2 and ECT 1.3. Based on the correlations observed between the ECT 1.2 and 1.3 and these three tests (Verbal Reasoning, Long-Term Memory, and Short-Term Memory), the following was observed. Firstly, the strongest relationships are consistent across test versions. Secondly, the correlations observed in the ECT 1.2 were higher than those observed in the ECT 1.3. Thirdly, for the ECT 1.2 and 1.3, the strongest correlations observed were between Verbal Reasoning and Long-Term Memory. Fourthly, the strongest relationship was observed between Long-Term Memory and Verbal Reasoning. Fifthly, the largest correlations were observed between Short-Term Memory and Verbal Reasoning, and Long-Term Memory (for the ECT 1.2) and Long-Term Memory (for the ECT 1.3).

Table 48: Correlations with Non-Verbal Tests for the ECT version 1.2 and 1.3

Psychometric Tests	**ECT 1.2**	**Non-Verbal Reasoning AAT 2**	**Numeric Comprehension AAT 5**	**ECT 1.3**	**Non-Verbal Reasoning AAT 2**	**Numeric Comprehension AAT 5**
ECT 1.2	**.79**	.392**	.427**	**.79**	.269**	.309**
Non-Verbal Reasoning	.392**	**.87**	.639**	.269**	**.87**	.557**
Verbal Reasoning	.640**	.659**	.694**	.410**	.595**	.535**
Vocabulary	.729**	.300**	.395**	.633**	.320**	.362**
Reading Comprehension	.711**	.478**	.489**	.637**	.311**	.339**
Numeric Comprehension	.427**	.639**	**.94**	.309**	.557**	**.94**

*p <.05, **p < .01, ***p < .001

In Table 48 above, the different psychometric tests used during the piloting of the ECT 1.2 and ECT 1.3 are indicated. These psychometric tests are not theorised to be related to the constructs of the ECT 1.2 and ECT 1.3. Based on the correlations observed both between the ECT 1.2 and 1.3 and these two tests (Non-Verbal Reasoning and Numeric Comprehension), the following was observed. Firstly, the correlations observed for these tests and the ECT across test versions were similar, but became smaller. Secondly, the correlations observed in the ECT 1.2 were higher than those observed in the ECT 1.3. Thirdly, for the ECT 1.2 and 1.3, the strongest correlations observed were between Non-Verbal Reasoning, Verbal Reasoning, and Numeric Comprehension. One would expect the relationship between Non-Verbal Reasoning and Numeric Comprehension to be the strongest. The relationship between Verbal Reasoning, Non-Verbal Reasoning, and Numeric Comprehension exists due to the commonality of reasoning present in these tests.

When comparing the mono-trait mono-method triangles across test versions (Figure 29 and Figure 30), there are a few observations that need to be noted. Firstly, the verbal reasoning triangle (which is comprised of the correlation between the ECT and the DAT and AAT tests as well as the reliability of the ECT) indicated that there were substantial relationships between constructs as the correlations ranged from moderate to large. Secondly, the correlations were lower within these mono-trait mono-method triangles for the ECT version 1.3. It should be noted that the verbal reasoning and reading-vocabulary triangles consistently indicated a strong relationship, while the memory triangle had a slightly smaller relationship emerging. These triangles also indicate the most related constructs for the ECT and as a result, emphasise the mutual connection among the constructs.

7.6 Differential Test Functioning Analysis Results

The differential test functioning (DTF) analyses results were obtained by following the procedures indicated in Chapter 6, section 6.7.2.4.2. The output for the DTF analyses will be divided into the two test versions; the ECT version 1.2 and ECT version 1.3. The specific focus of the DTF within these two test versions will be based on gender differences (males and females) and race differences (White and African; Coloured and African).

The output for the DTF for the ECT version 1.2 and ECT version 1.3 will be as follows: the average fit statistics for males and females, the DTF scatterplot with empirical trend line for gender, the DTF scatterplot with identity trend line for gender, DTF statistics for gender comparison, the average fit statistics for the African, White and Coloured racial groups, the DTF scatterplot with empirical trend line for the African and White racial groups, the DTF scatterplot with identity trend line for the African and White racial groups, DTF statistics for the African and White race groups, the DTF scatterplot with empirical trend line for the African and Coloured racial groups, and the DTF scatterplot with identity trend line for the African and Coloured racial groups and DTF statistics for the African and Coloured race groups.

7.6.1 The ECT Version 1.2 Results

7.6.1.1 Gender differences

Table 49: Average Fit Statistics for Male and Female Samples for the ECT version 1.2

Rasch Statistics	**Males**	**Females**
Total	395	158
Person: Infit MNSQ	1.01	1.01
Person: Outfit MNSQ	1.01	1.02
Person: Separation	1.84	2.00
Person: Reliability	0.77	0.80
Items: Infit MNSQ	0.99	0.99
Items: Outfit MNSQ	1.02	1.02
Items: Separation	9.72	6.14
Items: Reliability	0.99	0.97

In terms of Table 49 above, the males outnumber the females by almost two thirds. The person infit and outfit MNSQ values for males and females were both acceptable, as they are just over 1 (Linacre, 2002; Smith, Schumacker, & Bush, 1998). The items infit and outfit MNSQ values for males and females are similar and are considered acceptable. The person separation values for the males and females are below 2 (Baghaei & Amrahi, 2011) and suggests that there is limited variation in the abilities of both males and females.

The item separation values for both genders are, however, above 2 (Baghaei & Amrahi, 2011) and suggests there is some variation in the difficulties of the items in the test (there is more variation for the male sample than the female sample). The person reliability for the females (.80) is higher and more acceptable than the males. The item reliability for the males is higher than the females, but both are very good.

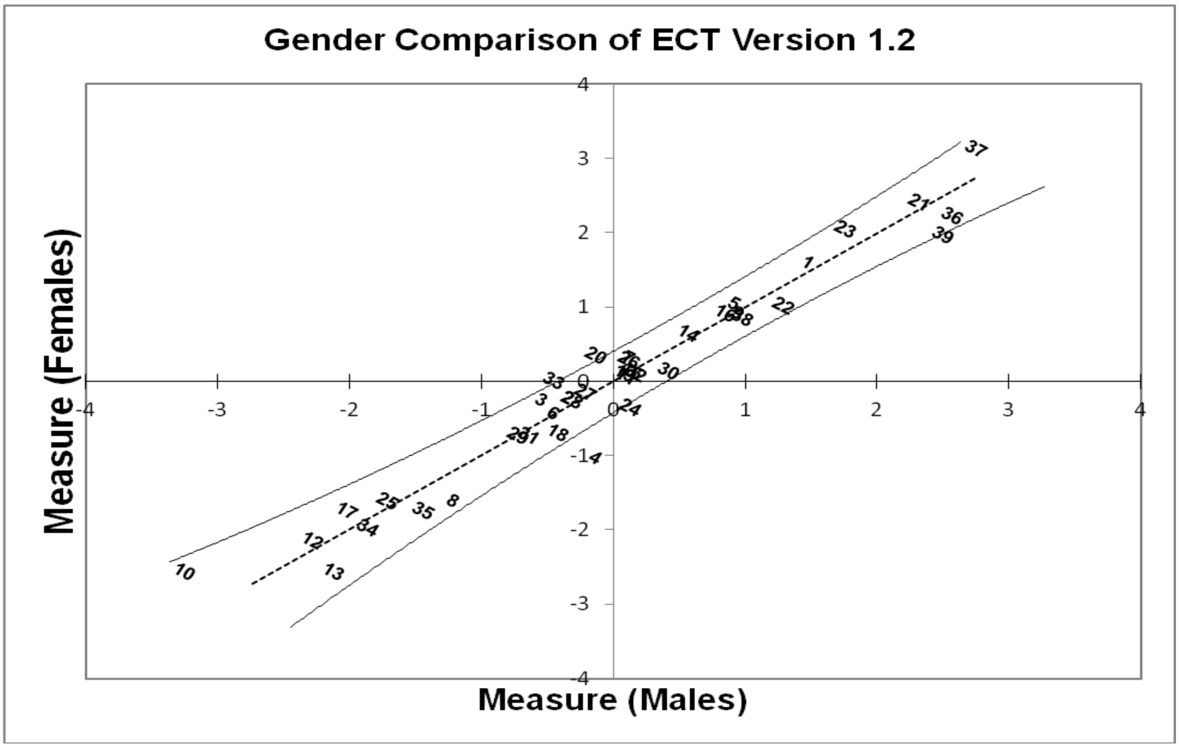


Figure 31: DTF Scatterplot for Gender with Empirical Line for the ECT version 1.2

The empirical trend line allows one to assess the best fitting linear connection between two sets of data (Linacre, 2009). In the DTF graph (Figure 31) with the empirical line, the only items that fall outside of the 95% confidence lines are items 4, 20, 35, and 24. These are the only items that can be considered as possibly biased for the gender group. Items 4 and 24 are considered easier for males than females. Items 20 and 35 are considered easier for females than males. The remaining items are clustered on either side of the empirical trend line, yet there seems to be slightly more items on the one side of the line. It is, however, not enough to suggest that one gender was favoured above the other. There are only a few outliers in terms of how the items are clustering.

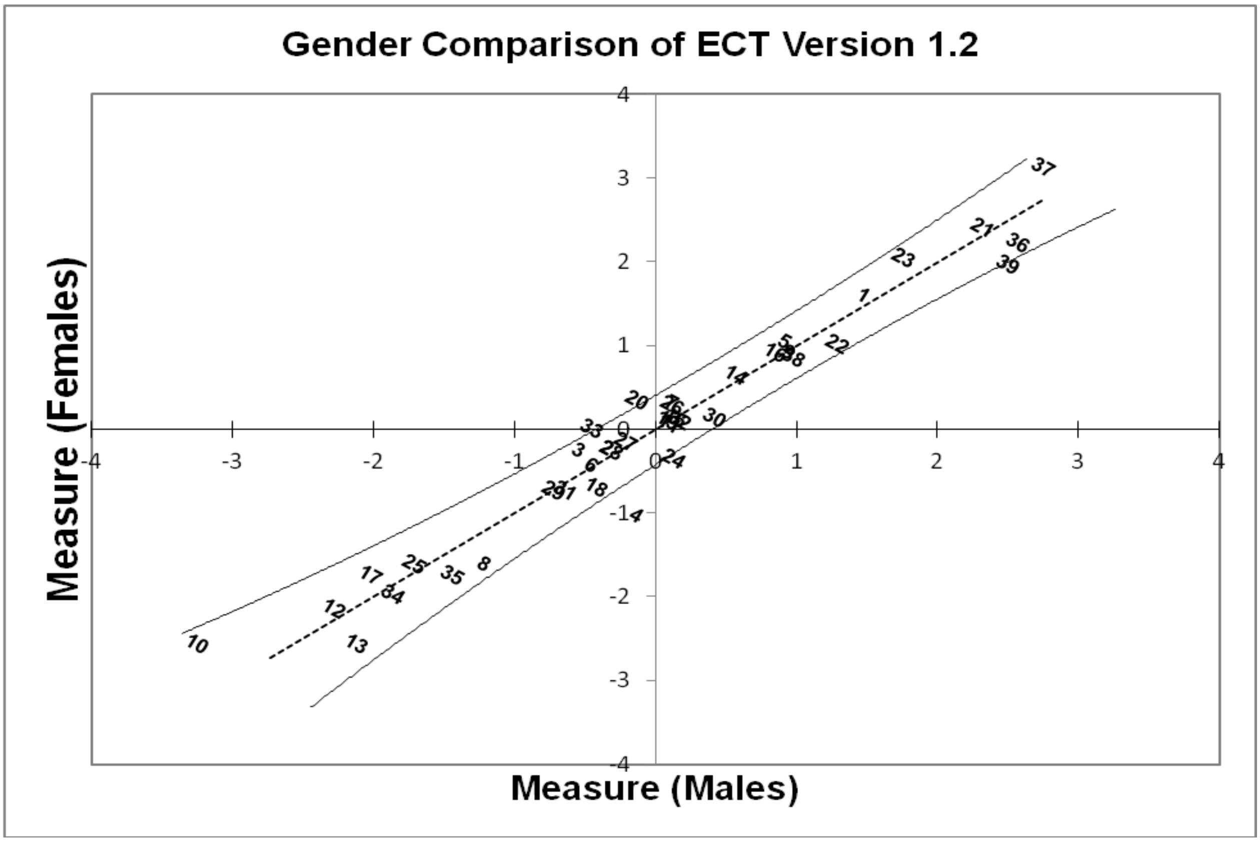


Figure 32: DTF Scatterplot for Gender with Identity Line for the ECT version 1.2

The identity trend line applies when it is hypothesised that two sets of data should be statistically similar; this line passes through the origin (Linacre, 2009). In the DTF graph with the identity line (Figure 32), the graph appears relatively similar to the empirical line. Most of the items lie along the identity line and form the mean. The items are distributed similarly on either side of the line. Items 4, 20, 24, and 35 are potentially DIF items. Items 4 and 24 are considered easier for males than females, while items 20 and 35 are considered easier for females than males. Items 36 and 39 border on the 95% confidence lines, but are within the confidence intervals. These items cannot be considered DIF but are possibly less comparable for the two genders. There are still a few outliers that are further apart from the other items. The *z*-scores of these items would need to be checked to determine the extent to which they depart from the 95% confidence lines.

Table 50: DTF Statistics for the Gender Comparison for the ECT version 1.2

ECT Item	Male Measure	Male SE	Female Measure	Female SE	*t* statistic
1	1.48	0.12	1.61	0.19	-0.58077
2	-0.34	0.12	-0.21	0.18	-0.6033
3	-0.54	0.12	-0.24	0.18	-1.38912
4	-0.13	0.11	-1.02	0.21	3.752075
5	0.92	0.11	1.06	0.18	-0.66609
6	-0.45	0.12	-0.42	0.19	-0.13578
7	0.12	0.11	0.33	0.17	-1.03965
8	-1.21	0.14	-1.6	0.25	1.35932
9	0.95	0.11	0.93	0.17	0.09624
10	-3.25	0.31	-2.55	0.35	-1.49827
11	0.1	0.11	0.08	0.18	0.092378
12	-2.28	0.2	-2.13	0.3	-0.41745
13	-2.12	0.19	-2.55	0.35	1.078447
14	0.57	0.11	0.66	0.17	-0.44701
15	0.14	0.11	0.15	0.18	-0.04984
16	0.85	0.11	0.93	0.17	-0.39762
17	-2.02	0.18	-1.73	0.26	-0.91868
18	-0.42	0.12	-0.68	0.2	1.112542
19	0.09	0.11	0.12	0.18	-0.14464
20	-0.13	0.11	0.36	0.17	-2.42247
21	2.32	0.14	2.42	0.22	-0.38545
22	1.29	0.11	1.03	0.18	1.230087
23	1.76	0.12	2.06	0.2	-1.28844
24	0.13	0.11	-0.34	0.19	2.138456
25	-1.71	0.17	-1.6	0.25	-0.36554
26	0.12	0.11	0.3	0.17	-0.89149
27	-0.21	0.11	-0.14	0.18	-0.33426
28	-0.31	0.12	-0.24	0.18	-0.32595
29	-0.72	0.12	-0.72	0.2	-0.0022
30	0.42	0.11	0.15	0.18	1.277492
31	-0.64	0.12	-0.72	0.2	0.340798
32	0.17	0.11	0.12	0.18	0.234592
33	-0.45	0.12	0.02	0.18	-2.17495
34	-1.86	0.17	-1.96	0.28	0.303716
35	-1.44	0.15	-1.73	0.26	0.964422
36	2.57	0.15	2.24	0.21	1.276737
37	2.75	0.16	3.14	0.28	-1.21093
38	0.98	0.11	0.87	0.17	0.540719
39	2.5	0.15	1.98	0.2	2.077949

Table 50 above lists the items and their corresponding measure values, standard errors, and the relevant *t* statistics. The *t* statistic is calculated as the difference between measures relative to their means (Linacre, 2012d). According to the DTF statistics table, the items that have the highest *t* statistic values (as they are greater than the 1.96 cut-off for 95% confidence) are items 4 (3.75), item 20 (-2.42), item 24 (2.13), item 33 (-2.17), and item 39 (2.07). These items were also identified in the DTF scatterplot as the items that lay on both the empirical and identity trend lines.

Table 51: DTF Statistics for the Male and Female Groups for the ECT version 1.2

Statistics	**Males**	**Females**
Mean	-1.1E-17	-0.00051
S.D.	1.369494	1.363493
Identity trend	2.732986	2.732474
Identity trend	-2.73299	-2.7335
Empirical trend	2.738987	2.726473
Empirical trend	-2.73899	-2.7275
Identity intercept	0.000513	
Identity slope	1	
Empirical intercept	-0.00051	0.000515
Empirical slope	0.995618	
Correlation	0.975147	
Reliability	0.975972	0.990084
Disattenuated correlation	0.992009	

In Table 51, the empirical slope of 0.996 is close to 1, which is considered acceptable. The empirical and identity slopes are close in range, which is why the items appeared very similar in the two scatterplots. This suggests that the items are similar for both genders. It is considered good if the empirical and identity slopes have similar values or intercepts. The correlation of the males and females is .975, which indicates that the items for both genders are measuring the same construct. The strong correlation suggests that the best fitting items are close to the origin. Their relationship is very strong (Cohen, 1988) and suggests that they are essentially equivalent. The reliability for both genders is well over .90 and indicates that there is high internal consistency present (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009). The disattenuated correlation is almost a perfect correlation, as this correlation is calculated without measurement error. The disattenuated correlation will consequently always be higher than the correlation.

7.6.1.2 White and African group differences

Table 52: Average Fit Statistics for African, White, and Coloured Samples for the ECT version 1.2

Rasch Statistics	**African**	**White**	**Coloured**
Total	428	71	44
Person: Infit MNSQ	1.01	1.01	1.01
Person: Outfit MNSQ	0.98	0.99	0.98
Person: Separation	1.70	1.56	1.69
Person: Reliability	0.74	0.71	0.74
Items: Infit MNSQ	0.99	1.00	0.99
Items: Outfit MNSQ	0.98	0.99	0.98
Items: Separation	10.87	3.12	2.78
Items: Reliability	0.99	0.91	0.89

In Table 52, the racial comparison consisted of the African, White, and Coloured groups. The African group is the reference group since it forms the majority of the sample. The Coloured group is the smallest and contains fewer individuals in comparison to the reference group.

The person infit and outfit MNSQ values for all racial groups were acceptable as most were 1 or very close to 1 (Linacre, 2002; Smith et al., 1998). The person separation values for all the race groups are relatively small (are under 2), which indicates that the range of person ability is very limited (Baghaei & Amrahi, 2011). The person reliability values for the three racial groups are all in the .7 range. This is a relatively low reliability value in terms of CTT (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009).

The item infit and outfit MNSQ values are all acceptable, since they are 1 or very close to 1 (Linacre, 2002; Smith et al., 1998). The item separation values for the White and Coloured groups are acceptable since they are above 2 (Baghaei & Amrahi, 2011), but are still considered small and imply that the range of item difficulty in the test is limited. The item separation value is however larger for the African group and suggests that there is a better spread of item difficulty for the African sample. The reliability values for the African and White groups are over .90 which is considered excellent. The Coloured group's reliability is just below .90, and is considered to be a good value (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009).

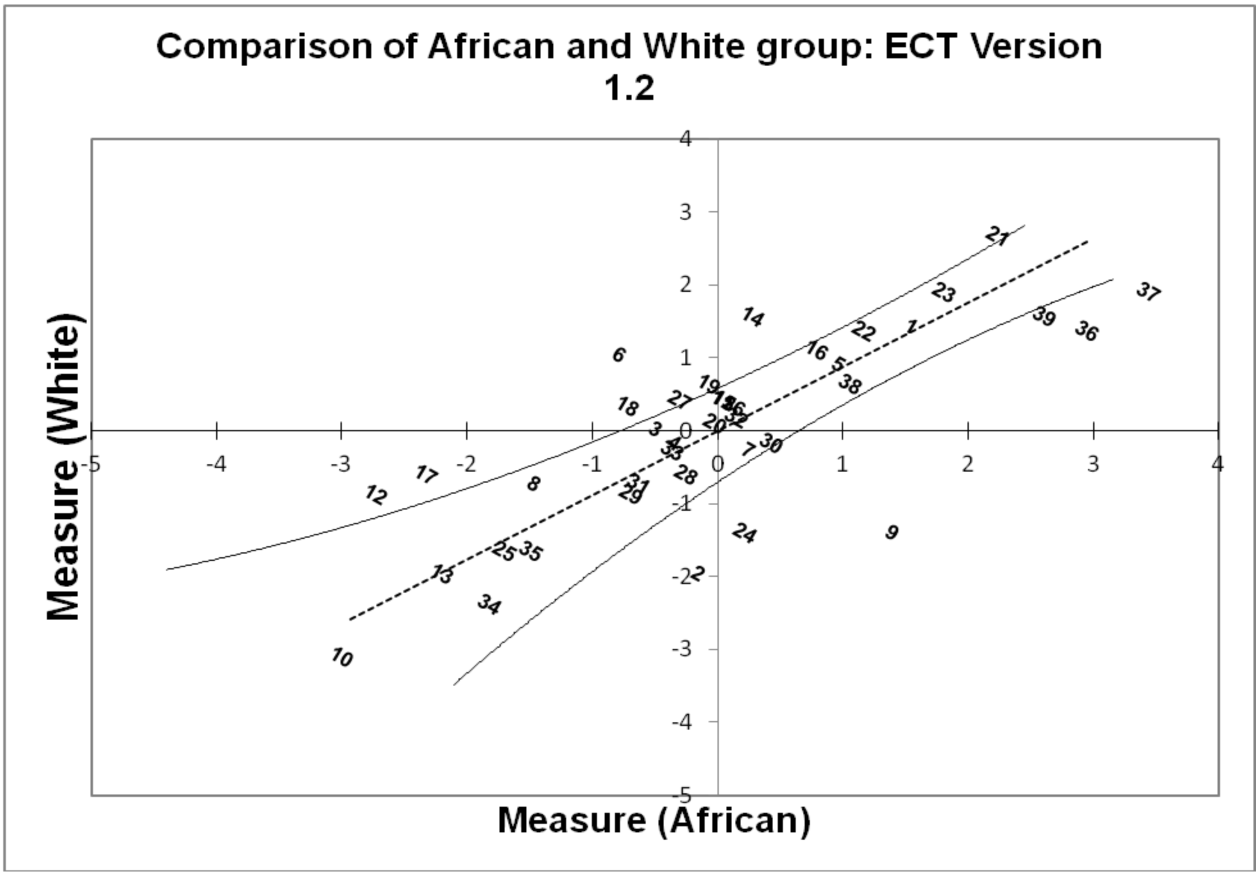

Comparison of African and White group: ECT Version 1.2

Figure 33: DTF Scatterplot for the African and White Groups with the Empirical Line for the ECT version 1.2

In the scatterplot of the African and White groups (Figure 33), the empirical line is observed. There are a few items that fall outside of the 95% confidence lines. These are items 37, 36, 39, 9, 24, 2, 12, 17, 18, 6, 27, 19, and 14. These items are potentially biased as they do not fall within the confidence intervals. Items 12, 17, 18, 6, 14, 27, and 19 can be considered to favour the White individuals over the African individuals. Items 9, 24, and 2 can be considered to favour the African individuals over the White individuals. Items 39, 36, 37, and 21 can be considered very difficult items for the African individuals. The items that border on the confidence lines are items 30 and 38. These items could potentially be biased. In terms of the item clustering across the line, the items seem to fall relatively equally on both sides of the line, even with the items falling outside of the confidence intervals.

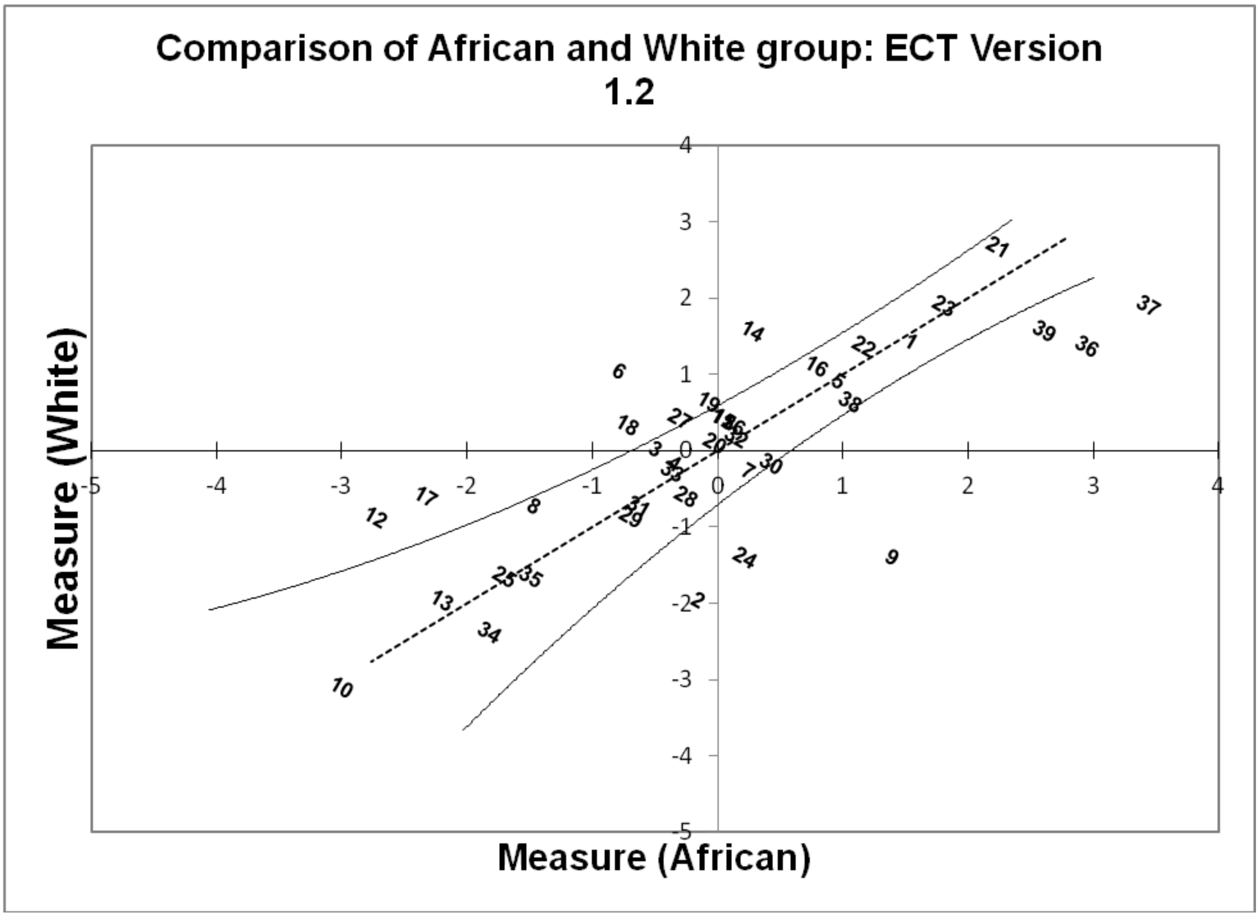


Figure 34: DTF Scatterplot of the African and White Group with an Identity Line for the ECT version 1.2

In the scatterplot of the African and White groups (Figure 34), the identity line is shown. There is evidence of numerous items that fall outside of the 95% confidence lines. These are items 37, 36, 39, 9, 24, 2, 12, 17, 18, 6, 27, 19, and 14. The items that border on the confidence lines are items 30 and 38.

In terms of which group is being favoured, items 6, 14, 18, 27, and 19 are more difficult for the White group. Items 39, 36, 37, 2, 24, and 9 are more difficult for the African group. Items 17 and 12 are easier for the White group. In terms of the item clustering across the line, the items seem to fall unequally on both sides of the line, including the items falling outside of the confidence intervals. This clustering suggests that the White group found more items easier than the African group. There are outliers observed, including those falling outside of the confidence intervals. The *z*-scores of these items would need to be checked to determine the extent to which they depart from the 95% confidence lines.

Table 53: DTF Statistics for the African and White Comparisons for the ECT version 1.2

ECT Items	**African Group Measure**	**African Group S.E.**	**White Group Measure**	**White Group S.E.**	***t* statistic**
1	1.55	0.12	1.44	0.26	0.380555
2	-0.15	0.11	-1.95	0.6	2.949138
3	-0.49	0.11	0.03	0.3	-1.6306
4	-0.34	0.11	-0.16	0.31	-0.55033
5	0.97	0.11	0.92	0.26	0.173476
6	-0.78	0.12	1.05	0.26	-6.39422
7	0.25	0.1	-0.26	0.32	1.518143
8	-1.46	0.14	-0.72	0.37	-1.87317
9	1.4	0.11	-1.39	0.47	5.777854
10	-3	0.24	-3.09	1.01	0.085707
11	0.04	0.1	0.43	0.27	-1.35809
12	-2.73	0.22	-0.86	0.38	-4.26114
13	-2.2	0.18	-1.95	0.6	-0.40073
14	0.28	0.1	1.58	0.26	-4.67041
15	0.05	0.1	0.43	0.27	-1.32336
16	0.79	0.1	1.12	0.26	-1.18831

17	-2.33	0.19	-0.59	0.35	-4.37173
18	-0.72	0.12	0.35	0.28	-3.51581
19	-0.07	0.11	0.65	0.27	-2.4731
20	-0.02	0.11	0.11	0.29	-0.42244
21	2.24	0.14	2.68	0.3	-1.33217
22	1.17	0.11	1.38	0.26	-0.74749
23	1.81	0.12	1.91	0.26	-0.3528
24	0.22	0.1	-1.39	0.47	3.348399
25	-1.7	0.15	-1.64	0.52	-0.11276
26	0.13	0.1	0.35	0.28	-0.74339
27	-0.3	0.11	0.43	0.27	-2.5074
28	-0.25	0.11	-0.59	0.35	0.923941
29	-0.69	0.11	-0.86	0.38	0.427134
30	0.43	0.1	-0.16	0.31	1.808168
31	-0.63	0.11	-0.72	0.37	0.2305
32	0.15	0.1	0.19	0.29	-0.13374
33	-0.36	0.11	-0.26	0.32	-0.29856
34	-1.82	0.15	-2.37	0.72	0.746438
35	-1.49	0.14	-1.64	0.52	0.276638
36	2.95	0.17	1.38	0.26	5.050707
37	3.44	0.21	1.91	0.26	4.574816
38	1.06	0.11	0.65	0.27	1.40277
39	2.61	0.15	1.58	0.26	3.428011

In the DTF statistics Table 53, the items that have the largest *t* statistic values (as they are greater than the 1.96 cut-off for 95% confidence) are items 2 (2.94), 6 (-6.39), 9 (5.77), 12 (-4.26), 14 (-4.67), 17 (-4.37), 18 (-3.51), 19 (-2.47), 24 (3.34), 27 (-2.50), 36 (5.05), 37 (4.57), and 39 (3.42). These items were all identified in the scatterplots (empirical and identity trend lines) as either outside of the 95% confidence lines or bordering on the confidence lines.

Table 54: DTF Statistics for the African and White Group Comparisons for the ECT version 1.2

Statistics	African Group	White Group
Mean	0.000256	-0.00077
S.D.	1.471271	1.297561
Identity trend	2.769088	2.768062
Identity trend	-2.76858	-2.7696
Empirical trend	2.942798	2.594352
Empirical trend	-2.94229	-2.59589
Identity intercept	0.001026	
Identity slope	1	
Empirical intercept	-0.001	0.001129
Empirical slope	0.881932	
Correlation	0.76503	
Reliability	0.911098	0.992027
Disattenuated correlation	0.8047	

In Table 54, the empirical slope is 0.881, which differs from the origin. The empirical trend line and identity trend are different to each other and indicate possible inequivalence across the groups. This was observed in the trend of the items across the line. There were slightly more items favouring the White group compared to the African group. The correlation between the African and White group is .76, which is not a satisfactory correlation because it suggests that the relationship between these two groups are not the same. It implies that the item locations are not equivalent across groups. The reliability values for both the African and White groups are very good as both are over .90 (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009), suggesting that there is a strong internal consistency present. The disattenuated correlation is .80 for the two groups, which is higher than the correlation observed. There is possibly another construct (20%) being measured between these two groups, because without measurement error, the disattenuated correlation could have been higher.

7.6.1.3 Coloured and African differences

The DTF scatterplot of the African and Coloured group with the empirical line is shown below (Figure 35). The items that fall outside of the 95% confidence lines are items 24, 9, 12, 18, 8, 19, 14, 38, 17, and 6. These items are potentially biased as they are favouring particular groups. Items 24, 9, and 38 are favouring the African group, which means that the African group performed better on these items than the Coloured group. Items 12, 17, 8, 6, 18, 19, and 14 are favouring the Coloured group, which means that the Coloured group performed better on these items than the African group. In terms of the item clustering, most of the items seem to cluster near the origin.

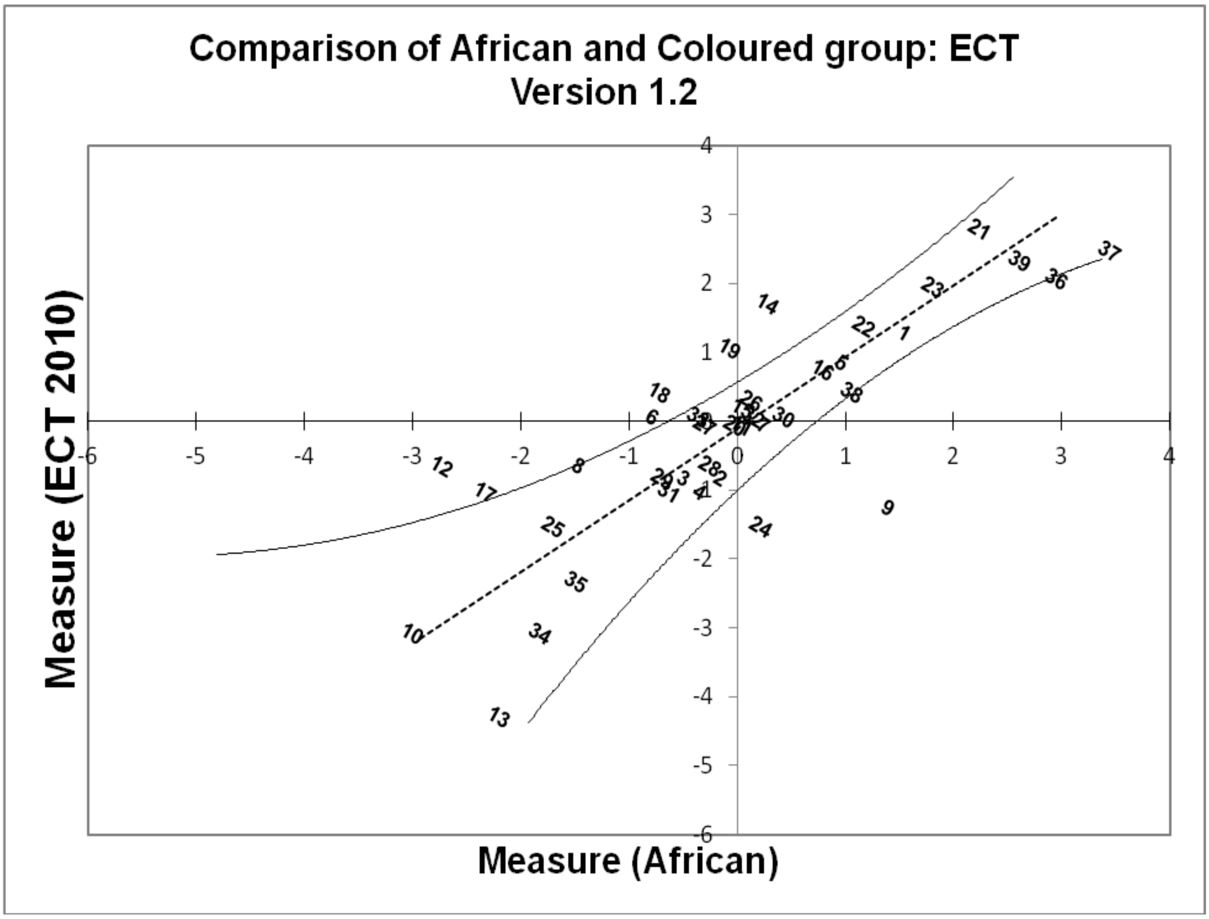


Figure 35: DTF Scatterplot for the African and Coloured Group with Empirical Line for the ECT version 1.2

In the DTF scatterplot for the African and Coloured groups (Figure 36), the identity is displayed. The items that fall outside of the 95% confidence lines are items 14, 19, 18, 6, 8, 12, 24, 9, 17, 36, and 38. These items are potentially biased as they are favouring a particular group. In terms of how the items are favouring the two groups, items 14, 19, 18, and 6 are more difficult for the Coloured group than the African group. Items 17, 8, 38, 24, 9, and 12 are easier for the Coloured group than the African group. In terms of the item clustering, most of the items seem to cluster near the origin. There are outliers observed, including those items outside of the confidence intervals. The *z*-scores of these items would need to be checked to determine the extent to which they depart from the 95% confidence lines.

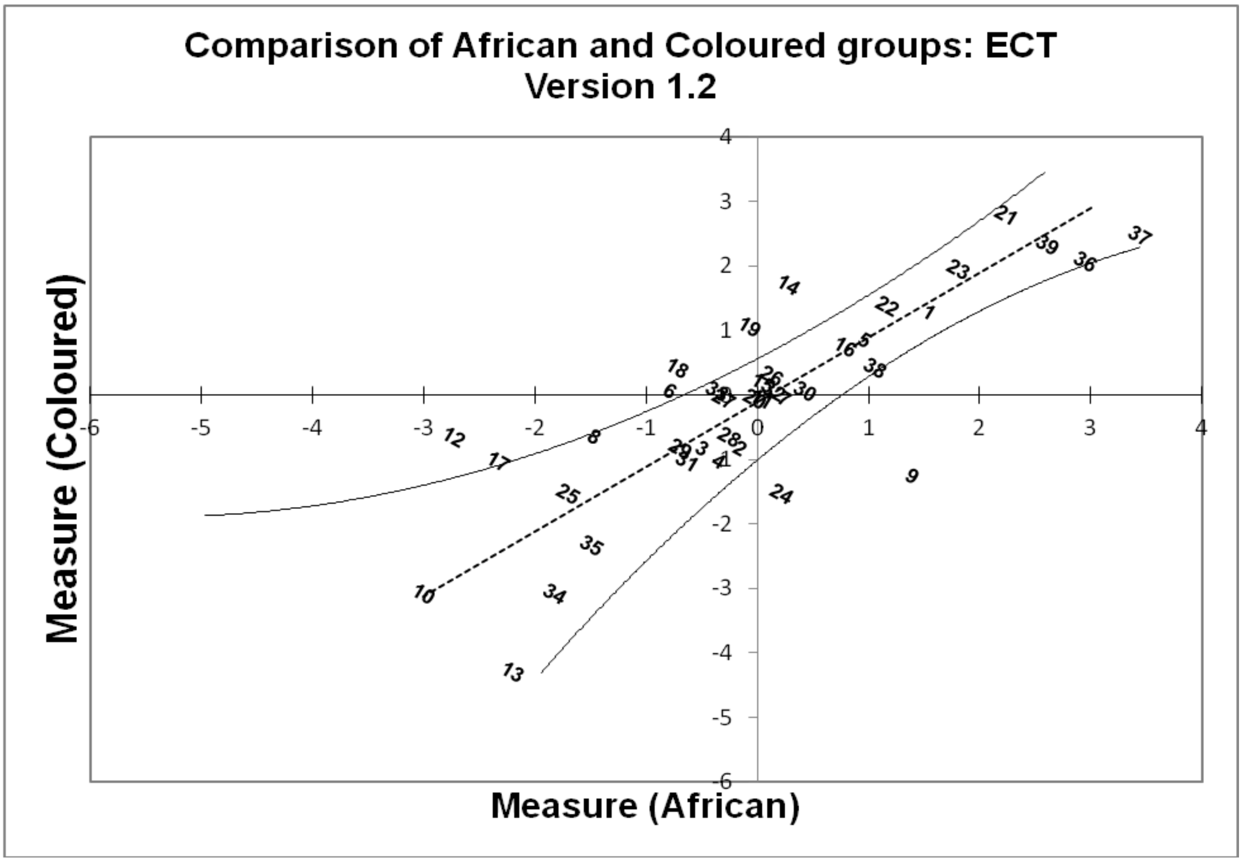


Figure 36: DTF Scatterplot for the African and Coloured Group with Identity Line for the ECT version 1.2

Table 55: DTF Statistics for the African and Coloured Group Comparisons for the ECT version 1.2

ECT Items	African Group Measure	African Group S.E.	Coloured Group Measure	Coloured Group S.E.	*t* statistic
1	1.55	0.12	1.29	0.33	0.427909
2	-0.15	0.11	-0.82	0.43	1.262274
3	-0.49	0.11	-0.82	0.43	0.496244
4	-0.34	0.11	-1.02	0.46	1.205694
5	0.97	0.11	0.86	0.33	0.000737
6	-0.78	0.12	0.07	0.35	-2.5939
7	0.25	0.1	-0.05	0.36	0.50921
8	-1.46	0.14	-0.64	0.41	-2.14601
9	1.4	0.11	-1.25	0.5	4.961855
10	-3	0.24	-3.07	1.03	-0.03758
11	0.04	0.1	-0.05	0.36	-0.05284
12	-2.73	0.22	-0.64	0.41	-4.72763
13	-2.2	0.18	-4.28	1.81	1.083196
14	0.28	0.1	1.72	0.34	-4.37285
15	0.05	0.1	0.2	0.35	-0.71357
16	0.79	0.1	0.76	0.33	-0.23126
17	-2.33	0.19	-1.02	0.46	-2.85264
18	-0.72	0.12	0.43	0.34	-3.4939
19	-0.07	0.11	1.07	0.33	-3.59276
20	-0.02	0.11	-0.05	0.36	-0.21184
21	2.24	0.14	2.8	0.41	-1.54588
22	1.17	0.11	1.39	0.33	-0.94795
23	1.81	0.12	1.96	0.35	-0.70201
24	0.22	0.1	-1.52	0.55	2.916291
25	-1.7	0.15	-1.52	0.55	-0.50824
26	0.13	0.1	0.31	0.34	-0.81756
27	-0.3	0.11	-0.05	0.36	-0.95567
28	-0.25	0.11	-0.64	0.41	0.660204
29	-0.69	0.11	-0.82	0.43	0.045638
30	0.43	0.1	0.07	0.35	0.687507
31	-0.63	0.11	-1.02	0.46	0.592547
32	0.15	0.1	0.07	0.35	-0.08171
33	-0.36	0.11	0.07	0.35	-1.47118
34	-1.82	0.15	-3.07	1.03	1.095489
35	-1.49	0.14	-2.32	0.74	0.956355
36	2.95	0.17	2.08	0.35	1.953877
37	3.44	0.21	2.48	0.38	1.958368
38	1.06	0.11	0.43	0.34	1.455868
39	2.61	0.15	2.34	0.37	0.401394

In the DTF statistics Table 55, the items that have the highest *t* statistic values (as they are greater than the 1.96 cut-off for 95% confidence) are items 6 (-2.59), 8 (-2.14), 9 (4.96), 12 (-4.72), 14 (-4.37), 17 (-2.85), 18 (-3.49), 19 (-3.59), and 24 (2.91). These items fell outside of the 95% confidence lines for both the empirical trend and identity trend line scatterplots.

Table 56: DTF Statistics for the African and Coloured Group Comparisons for the ECT version 1.2

Statistics	African Group	Coloured Group
Mean	0.000256	-0.10949
S.D.	1.471271	1.53439
Identity trend	3.005918	2.896174
Identity trend	-3.0054	-3.11515
Empirical trend	2.942798	2.959293
Empirical trend	-2.94229	-3.17827
Identity intercept	0.109744	
Identity slope	1	
Empirical intercept	-0.10975	0.10524
Empirical slope	1.042901	
Correlation	0.814255	
Reliability	0.879807	0.992027
Disattenuated correlation	0.871576	

In Table 56, the empirical slope is 1.04, which is close to the origin. This does not however make the empirical and identity slopes identical as there are differences across groups. The similarity between the empirical and identity trend lines is evident in the scatterplots and in the trend of the items. The correlation between the African and Coloured groups is .81, which is not a satisfactory correlation as it suggests that the item locations are not equivalent for these two groups. The reliability value for the African group is excellent (.99), while the Coloured group has a slightly lower, although still good, reliability (.88) (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009). The high reliabilities suggest that there is a good internal consistency present for these two groups. The disattenuated correlation is .87, which is good, but suggests that there may be another construct (13%) being measured as this correlation should have been higher if measurement error is eliminated.

7.6.2 The ECT Version 1.3 Results

7.6.2.1 Gender differences

Table 57: Average Fit Statistics for Male and Female Samples for the ECT version 1.3

Rasch Statistics	Males	Females
Total	665	212
Person: Infit MNSQ	1.00	1.00
Person: Outfit MNSQ	1.02	1.02
Person: Separation	1.81	1.80
Person: Reliability	0.77	0.76
Items: Infit MNSQ	0.99	0.98
Items: Outfit MNSQ	1.02	1.02
Items: Separation	13.60	7.77
Items: Reliability	0.99	0.98

In Table 57 above, there are fewer females compared to the males. The person infit and outfit MNSQ values for males and females were both acceptable, as they are close to and equal to 1 (Linacre, 2002; Smith et al., 1998). The item infit and outfit MNSQ values for males and females are acceptable (both are equal to 1). The person separation values for the males and females are below 2 (Baghaei & Amrahi, 2011) and suggests that there is limited variation in the abilities of males and females. The item separation value is higher than 2 (Baghaei & Amrahi, 2011) which suggests that there is a range of item difficulties in the test. It is clear from the item separation values that there is substantially more variation of item difficulties for the male sample than the female sample. The person reliability for the males (.77) and females (.76) are relatively similar to each other, and are considered to be poor reliabilities. The item reliability for the male sample is higher than the female sample, but both are excellent reliability values.

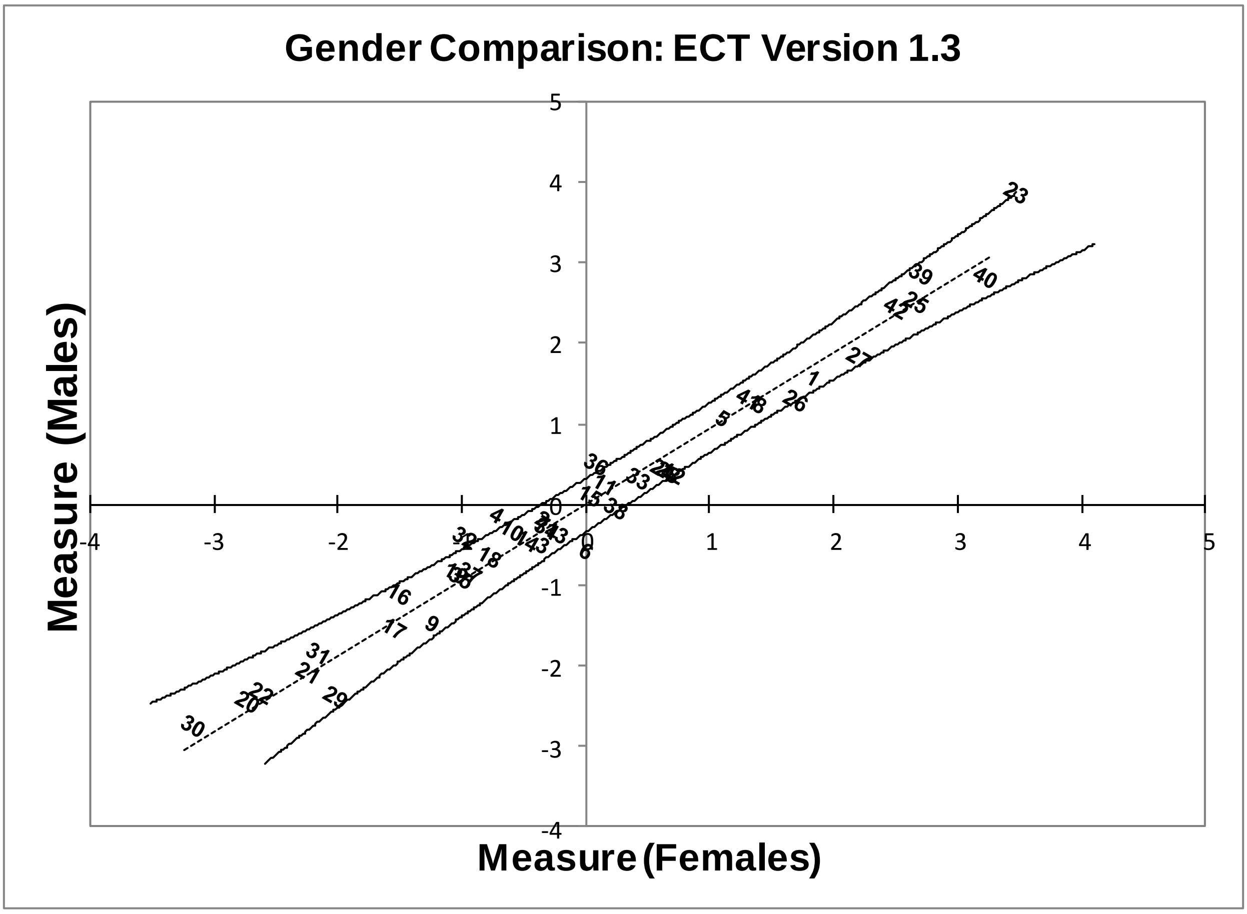


Figure 37: DTF Scatterplot for Gender with Empirical Line for the ECT version 1.3

In the DTF graph with the empirical line (Figure 37), the only items that fall outside of the 95% confidence lines are items 23, 36, 6, 4, and 32. These are the only items that can be considered as possibly biased for the gender group. Items 38, 12, 24, 26, and 28 are bordering on the confidence intervals, which means that they could also possibly be considered biased. Items 36, 32, and 4 are considered easier for males than females. Items 6 and 38 can be considered easier for females than males. Item 23 could be considered difficult for both genders. The remaining items are clustered on either side of the empircal trend line, yet there seems to be slightly more items on the one side of the line.

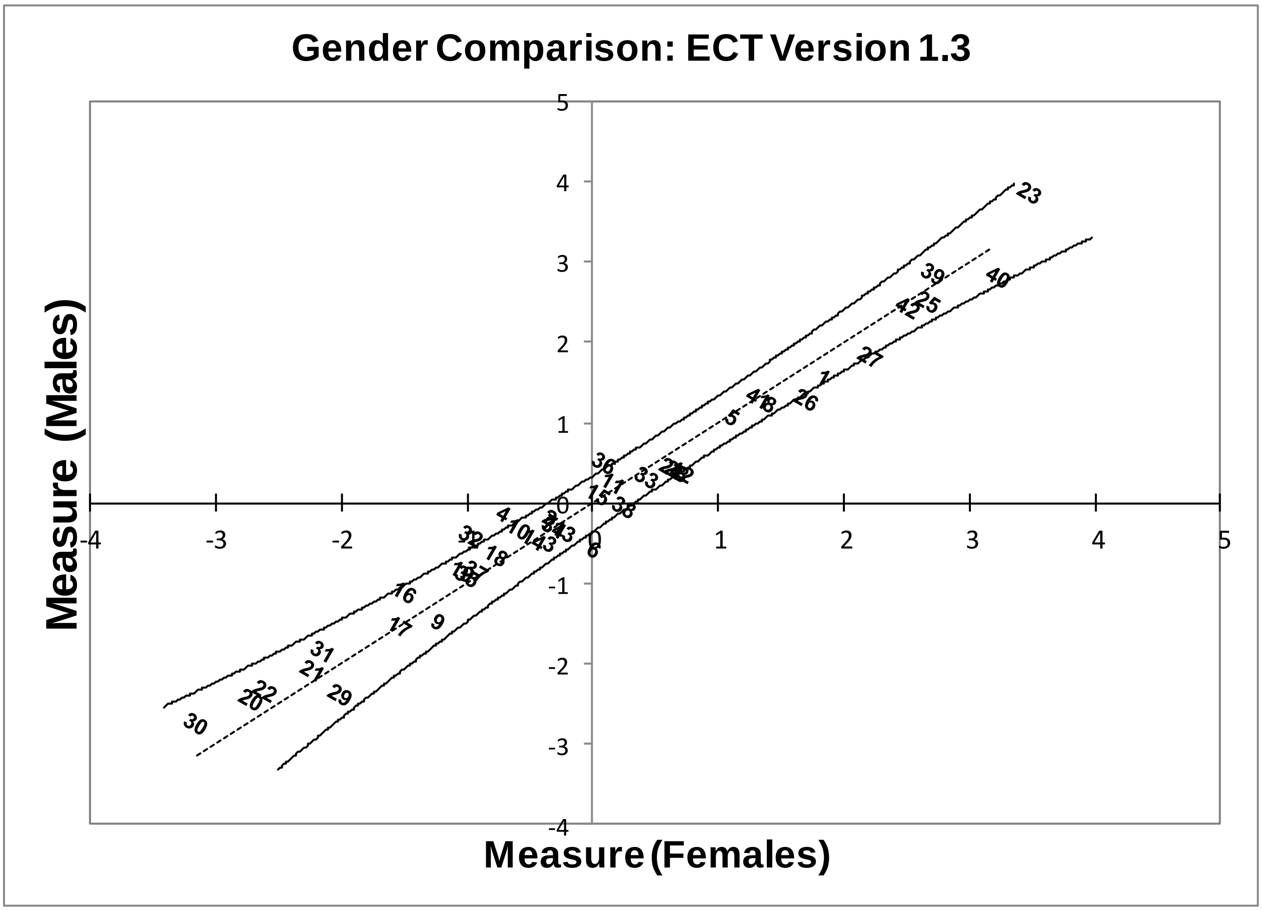


Figure 38: DTF Scatterplot for Gender with Identity Line for the ECT version 1.3

In the DTF graph with the identity line (Figure 38), the graph appears relatively similar to the empirical line. Items 4, 32, 36, 38, 27, and 26 are potentially DIF items. Items 4, 32, and 36 are easier for males than females, while items 6, 12, 24, and 38 are easier for females than males. Items 27 and 28 border on the 95% confidence lines, but are within the confidence intervals. These items cannot be considered DIF but are possibly less comparable for the two genders. There are still a few outliers that are further apart from the other items.

Table 58: DTF Statistics for the Gender Comparisons for the ECT version 1.3

ECT Item	Male Measure	Male S.E.	Female Measure	Female S.E.	*t* statistic
1	1.66	0.16	1.46	0.09	-1.09725
2	-0.49	0.16	-0.29	0.09	1.081687
3	-0.49	0.16	-0.61	0.09	-0.66146
4	-0.86	0.18	-0.24	0.09	3.073706
5	0.95	0.15	0.95	0.08	-0.0084
6	-0.14	0.15	-0.69	0.09	-3.15231
7	-0.49	0.16	-0.35	0.09	0.754847
8	1.24	0.15	1.11	0.09	-0.75133
9	-1.38	0.2	-1.59	0.12	-0.90649
10	-0.74	0.17	-0.42	0.09	1.656174
11	0	0.15	0.14	0.08	0.815126
12	0.55	0.15	0.28	0.08	-1.59664
13	-0.38	0.16	-0.46	0.09	-0.44357
14	-0.62	0.17	-0.56	0.09	0.304498
15	-0.12	0.15	0	0.09	0.677828
16	-1.65	0.22	-1.22	0.11	1.742391
17	-1.7	0.22	-1.65	0.12	0.193821
18	-0.92	0.18	-0.76	0.1	0.770091
19	-1.19	0.19	-0.96	0.1	1.064563
20	-2.88	0.35	-2.56	0.17	0.818736
21	-2.39	0.29	-2.2	0.15	0.57756
22	-2.77	0.33	-2.44	0.16	0.895919
23	3.28	0.26	3.74	0.18	1.45013
24	0.47	0.15	0.32	0.08	-0.89076
25	2.48	0.2	2.37	0.11	-0.48818
26	1.53	0.16	1.17	0.09	-1.96883
27	2.03	0.18	1.7	0.09	-1.64688
28	0.51	0.15	0.29	0.08	-1.30252
29	-2.16	0.26	-2.5	0.17	-1.0991
30	-3.32	0.42	-2.86	0.2	0.985776
31	-2.31	0.28	-1.95	0.14	1.145414
32	-1.12	0.19	-0.52	0.09	2.847114
33	0.27	0.15	0.22	0.08	-0.30252
34	-0.46	0.16	-0.41	0.09	0.264585
35	-1.15	0.19	-1.01	0.1	0.645391
36	-0.07	0.15	0.39	0.08	2.697479
37	-1.08	0.19	-0.96	0.1	0.552242
38	0.09	0.15	-0.16	0.09	-1.43732
39	4.61	0.46	4.19	0.21	-0.83341
40	4.84	0.51	3.65	0.17	-2.21625
41	2.56	0.2	2.23	0.1	-1.48219
42	3.77	0.31	3.18	0.14	-1.73874

According to the DTF statistics table (table 58), the items that have the highest *t* statistic values (as they are greater than the 1.96 cut-off for 95% confidence) are items 4 (-2.83235), 6 (3.429972), 26 (2.287886), 27 (1.987616), 32 (-2.61608), and 36 (-2.35294). This indicates that these items are statistically different for the two genders and can be considered biased. These items were also identified in the DIF scatterplots as the items that lay on both the empirical and identity trend lines.

Table 59: DTF Statistics for the Male and Female Groups for the ECT version 1.3

Statistics	**Males**	**Females**
Mean	-0.00095	0.000476
S.D.	1.914158	1.708632
Identity trend	3.621837	3.623266
Identity trend	-3.62374	-3.62231
Empirical trend	3.827363	3.41774
Empirical trend	-3.82927	-3.41679
Identity intercept	0.001429	
Identity slope	1	
Empirical intercept	-1.38687E-05	1.47265E-05
Empirical slope	0.941751	
Correlation	0.986065	
Reliability	0.994904151	0.984493321
Disattenuated correlation	0.996342012	

In Table 59, the empirical slope of 0.942 is slightly under 1, which is considered acceptable. The empirical slope and the identity slope are different in terms of their values, which is why the items appeared very similar in the two scatterplots; with the exception of a few items being considered DIF in one scatterplot and not the other. This suggests that the items are relatively similar for both genders, with only a few differences. This was confirmed by the *t* statistics. It is considered good if the empirical and identity slopes have similar values or intercepts. The correlation of the males and females is .986, which indicates that the items of both genders are measuring the same construct. The strong correlation suggests that the best fitting items are close to the origin. Their relationship is very strong (Cohen, 1988) and suggests that they are essentially equivalent. The reliability for both genders is well over .90 and indicates that there is high internal consistency present (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009). The disattenuated correlation is almost a perfect correlation, as this correlation is calculated without measurement error. The disattenuated correlation will consequently always be higher than the correlation.

7.6.2.2 White and African group differences

Table 60: Average Fit Statistics for the African, White, and Coloured Samples for the ECT version 1.3

Rasch Statistics	**African**	**White**	**Coloured**
Total	681	135	50
Person: Infit MNSQ	1.00	1.00	1.00
Person: Outfit MNSQ	1.03	0.93	1.06
Person: Separation	1.67	1.38	1.73
Person: Reliability	0.74	0.66	0.75
Items: Infit MNSQ	0.99	1.00	0.95
Items: Outfit MNSQ	1.03	0.93	1.08
Items: Separation	14.33	4.55	3.44
Items: Reliability	1.00	0.95	0.92

The ethnic comparison consisted of the African, White, and Coloured groups (Table 60). The Coloured group is the smallest group in comparison to the African and White group. The person infit and outfit MNSQ values for all racial groups were acceptable as most were 1 or very close to 1. The item infit and outfit MNSQ values are all acceptable, since they are 1 or very close to 1 (Linacre, 2002; Smith et al., 1998). The person separation values for all the ethnic groups are relatively small (all are under 2), which indicates that the range of person abilities across the different race groups are very limited. The item separation values for the White and Coloured groups are acceptable since they are above 2 (Baghaei & Amrahi, 2011), but are still considered small and imply that the range of item difficulties are limited in the test for these race groups. The item separation value for the African group (14.33) is large, suggesting that there is a greater range of item difficulties across the test for the African group. The person reliability values for the African and Coloured groups are in the .70 range and the White group has a .66 reliability. These are relatively low reliability values (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009). The item reliability values for the African, White, and Coloured groups are all over .90 and are excellent reliability values (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009).

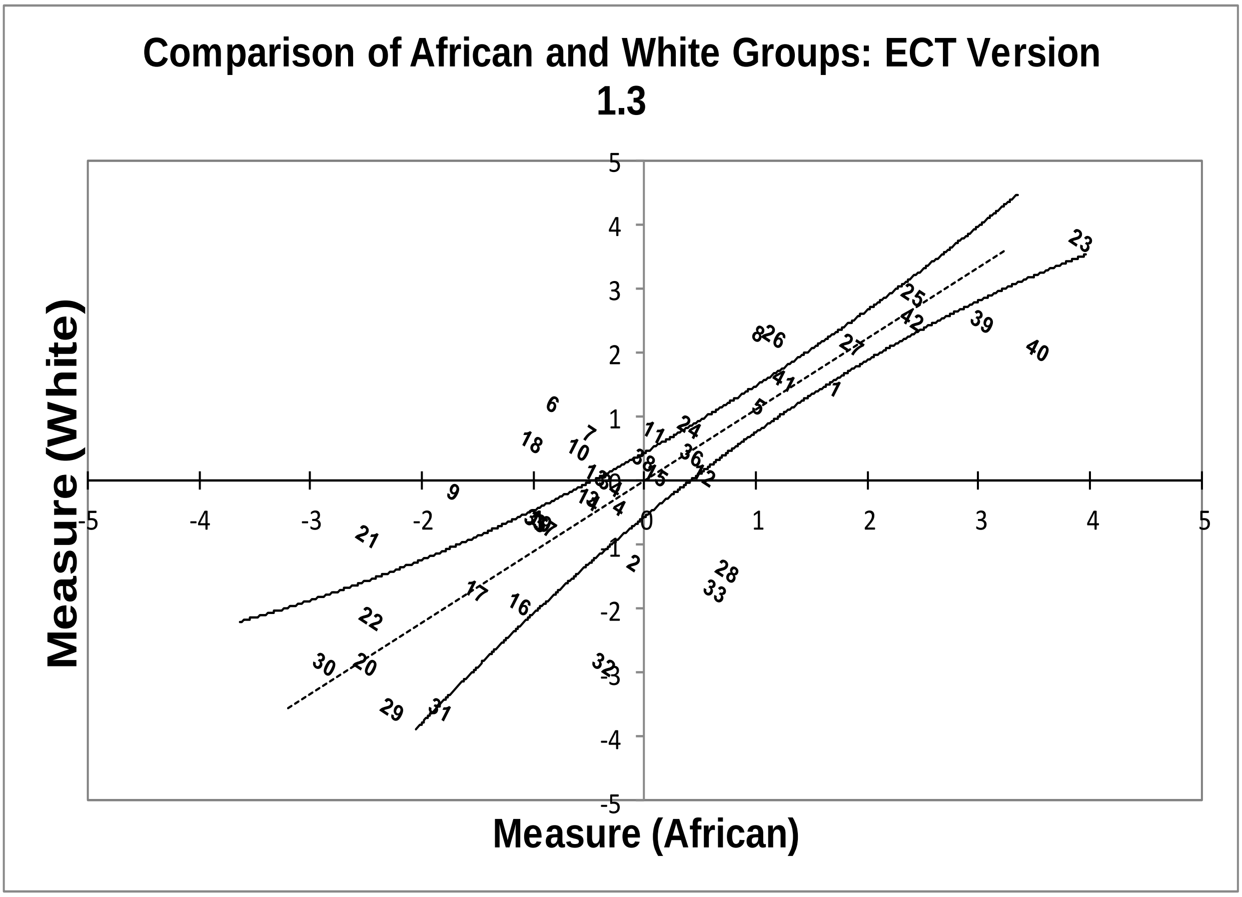


Figure 39: DTF Scatterplot for the African and White Groups with the Empirical Line for the ECT version 1.3

In the scatterplot of the African and White groups with the empirical line (Figure 39), there are a few items that fall outside of the 95% confidence lines. These are items 1, 2, 6, 7, 8, 9, 10, 11, 12, 18, 21, 26, 28, 31, 32, 33, 39, and 40. These items are potentially biased as they do not fall within the confidence intervals. Items 31, 13, 38 and 24 are bordering on the confidence intervals. Items 21, 9, 18, 10, 6, 7, 11, 13, 24, 8, and 26 can be considerd to favour the White individuals over the African individiuals. Items 28, 33, 32, 2, 12, 1, and 31 can be considered to favour the African individuals over the White individuals.

Items 39 and 40 can be considered very difficult items for the African and White individuals. These items could potentially be biased. In terms of the item clustering across the lines, the items seem to fall relatively equally on both sides of the line, even with the items falling outside of the confidence intervals.

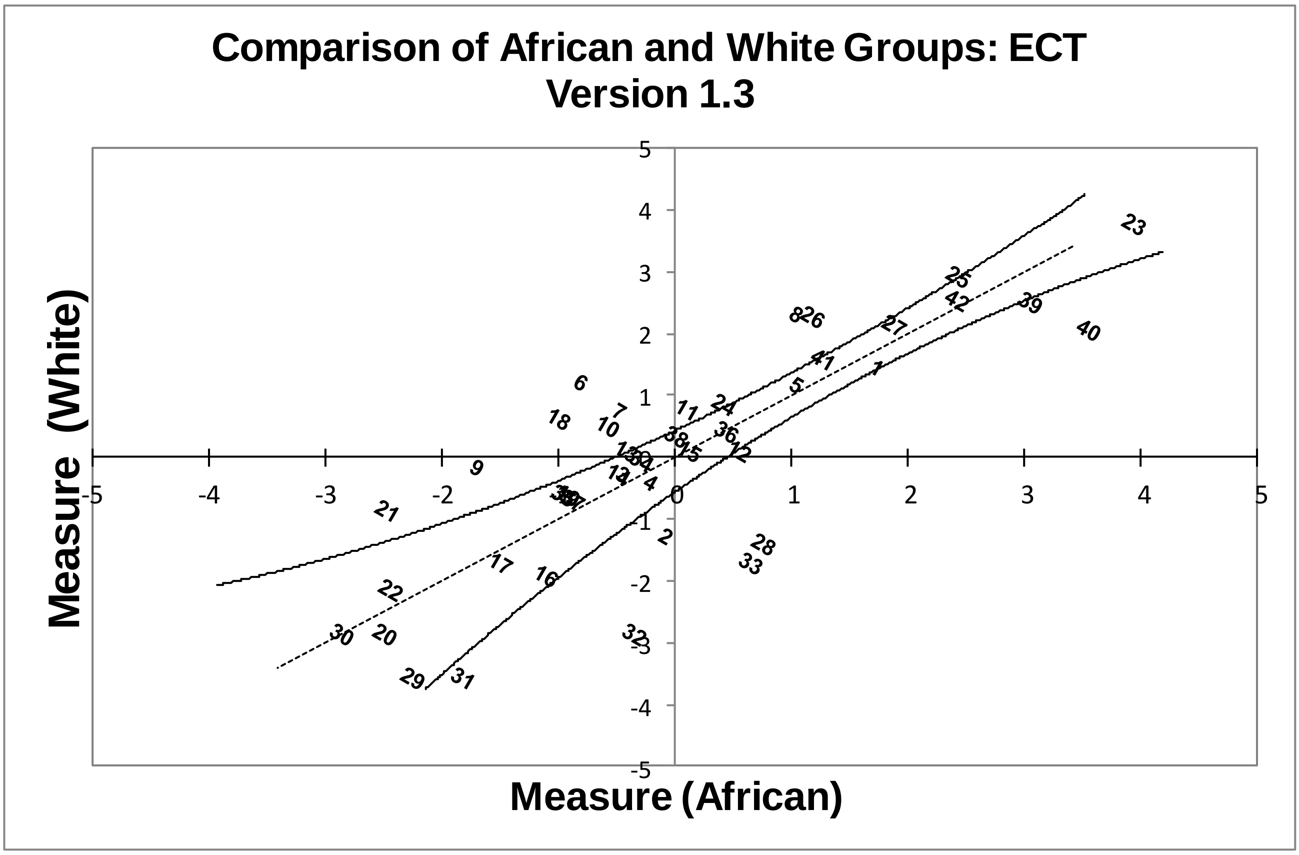


Figure 40: DTF Scatterplot for the African and White Groups with the Identity Line for the ECT version 1.3

In the scatterplot of the African and White groups with the identity line (Figure 40), there is evidence of numerous items that fall outside of the 95% confidence lines. These are items 2, 6, 7, 8, 9, 10, 11, 18, 21, 26, 28, 31, 32, 39, and 40. The items that border on the confidence lines are items 1, 12, 39, 25, 13, and 24.

In terms of which group is being favoured, items 8 and 26 are more difficult for the African group. Items 39 and 40 are more difficult for the White group. Items 21, 9, 18, 7, 10, 11, 24, and 6 are easier for the White group, while items 28, 33, 2, 32, 31, and 12 are easier for the African group. In terms of the item clustering across the line, the items seem to fall unequally on both sides of the line, including the items falling outside of the confidence intervals. This clustering suggests that the White group found more items easier than the African group.

Table 61: DTF Statistics for the African and White Comparisons for the ECT version 1.3

ECT Items	**African Group Measure**	**African Group S.E.**	**White Group Measure**	**White Group S.E.**	***t* statistic**
1	1.73	0.09	1.44	0.18	1.442205
2	-0.08	0.08	-1.29	0.35	3.370888
3	-0.45	0.09	-0.29	0.25	-0.60127
4	-0.21	0.09	-0.42	0.26	0.764123
5	1.04	0.08	1.16	0.19	-0.58093
6	-0.81	0.09	1.2	0.19	-9.55946
7	-0.48	0.09	0.73	0.2	-5.51604
8	1.04	0.08	2.3	0.19	-6.11074
9	-1.7	0.12	-0.18	0.24	-5.66382
10	-0.58	0.09	0.49	0.2	-4.87769
11	0.1	0.08	0.76	0.19	-3.20032
12	0.55	0.08	0.09	0.22	1.96604
13	-0.42	0.09	0.09	0.22	-2.14458
14	-0.49	0.09	-0.29	0.25	-0.75181
15	0.12	0.08	0.09	0.22	0.129171
16	-1.11	0.1	-1.93	0.46	1.742429
17	-1.5	0.11	-1.73	0.42	0.5303
18	-1	0.1	0.61	0.2	-7.19907

19	-0.92	0.1	-0.63	0.27	-1.00638
20	-2.49	0.16	-2.88	0.72	0.529091
21	-2.47	0.16	-0.88	0.3	-4.67577
22	-2.44	0.16	-2.16	0.51	-0.5234
23	3.94	0.2	3.76	0.27	0.536413
24	0.42	0.08	0.84	0.19	-2.03614
25	2.43	0.11	2.91	0.21	-2.02375
26	1.18	0.09	2.26	0.19	-5.1359
27	1.88	0.1	2.12	0.19	-1.11668
28	0.76	0.08	-1.42	0.37	5.759448
29	-2.25	0.15	-3.59	1.01	1.312572
30	-2.86	0.19	-2.88	0.72	0.027178
31	-1.82	0.13	-3.59	1.01	1.73837
32	-0.35	0.09	-2.88	0.72	3.487082
33	0.65	0.08	-1.73	0.42	5.567142
34	-0.29	0.09	-0.07	0.23	-0.88979
35	-0.96	0.1	-0.63	0.27	-1.14531
36	0.44	0.08	0.4	0.21	0.179057
37	-0.88	0.1	-0.71	0.28	-0.57097
38	0.01	0.08	0.32	0.21	-1.37842
39	3.05	0.14	2.49	0.2	2.294825
40	3.55	0.17	2.05	0.19	5.884418
41	1.27	0.09	1.57	0.18	-1.48953
42	2.42	0.11	2.53	0.2	-0.48088

In the DTF statistics Table 61, the items that have the largest *t* statistic values are items 2 (3.37), 6 (-9.56), 7 (-5.52), 8 (-6.11), 9 (-5.66), 10 (-4.88), 11 (-3.20), 13 (-2.14), 18 (-7.20), 21 (-4.68), 24 (-2.04), 25 (-2.02), 26 (-5.14), 28 (5.76), 32 (3.49), 33 (5.57), 39 (2.30), and 40 (5.89). These items were all identified in the scatterplots (empirical and identity trend lines) as either outside of the 95% confidence lines or bordering on the confidence lines.

Table 62: DTF Statistics for the African and White Group Comparisons for the ECT version 1.3

Statistics	**African Group**	**White Group**
Mean	0.000476	0.000714
S.D.	1.613967	1.796231
Identity trend	3.410674	3.410912

Identity trend	-3.40972	-3.40948
Empirical trend	3.228411	3.593176
Empirical trend	-3.22746	-3.59175
Identity intercept	-0.00024	
Identity slope	1	
Empirical intercept	-0.000165617	0.00018432
Empirical slope	1.112928848	
Correlation	0.818283065	
Reliability	0.995368231	0.955332179
Disattenuated correlation	0.839140094	

In Table 62, the empirical slope is 1.113, which differs from the origin. The empirical trend line and identity trend line are different and suggest there is inequivalence across the groups. This was observed in the trend of the items across the line. There were only a few items favouring the White group compared to the African group. The correlation between the African and White group is .818, which is not a satisfactory correlation because the item locations across the two groups are not equivalent. The reliability values for both the African and White groups are very good as both are over .90 (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009), suggesting that there is a strong internal consistency present. The disattenuated correlation obtained for these two groups is .839, which is higher than the correlation. The disattenuated correlation could however have been higher if it contained less measurement error. This allows one to consider that there is possibly another construct (16%) being measured between these two groups.

7.6.2.3 Coloured and African group differences

In the DTF scatterplot of the African and Coloured group on the empirical line (Figure 41), the items that fall outside of the 95% confidence lines are items 21, 19, 18, 6, 10, 7, 32, 33, 28, 23, and 3. These items are potentially biased as they are favouring particular groups. Items 38, 12, 24, and 26 are bordering on the 95% confidence intervals. Items 28, 32, 3, and 33 are favouring the African group, which means that the African group performed better on these items than the Coloured group. Items 6, 7, 9, 10, 18, and 21 are favouring the Coloured group, which means that the Coloured group performed better on these items than the African group. In terms of the item clustering, most of the items seem to cluster near the origin.

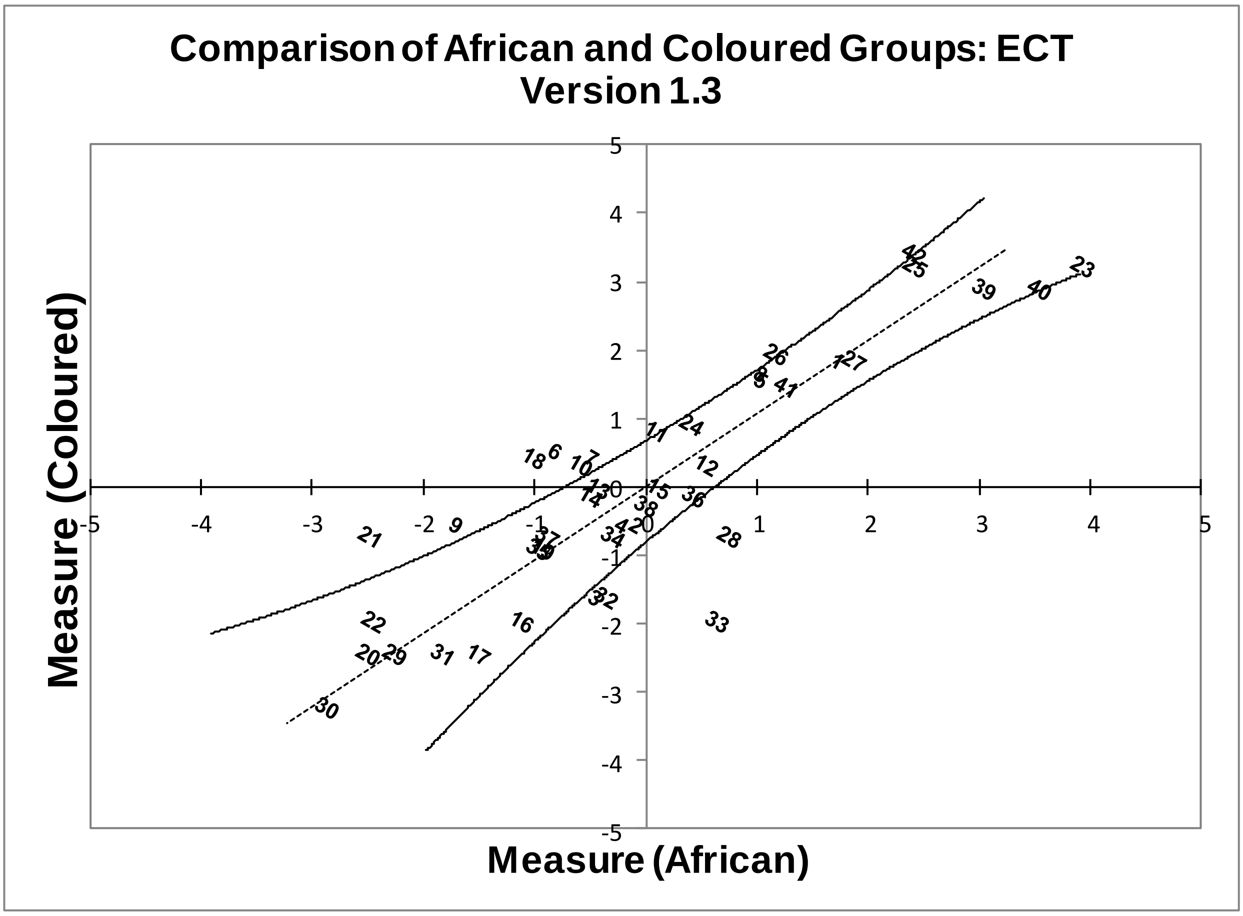


Figure 41: DTF Scatterplot for the African and Coloured Groups with Empirical Line for the ECT version 1.3

In the DTF scatterplot for the African and Coloured groups on the identity line (Figure 42), the items that fall outside of the 95% confidence lines are items 21, 9, 18, 6, 10, 7, 32, 3, 28, 33, and 42. The items that border on the 95% confidence lines are items 11, 26, 25, and 36. These items are potentially biased as they are favouring a particular group. In terms of how the items are favouring the two groups, items 28, 33, 32, and 3 are more difficult for the Coloured group than the African group. Items 21, 9, 18, 6, 7, and 10 are easier for the Coloured group than the African group. Item 42 is difficult for both groups. In terms of the item clustering, most of the items seem to cluster near the origin.

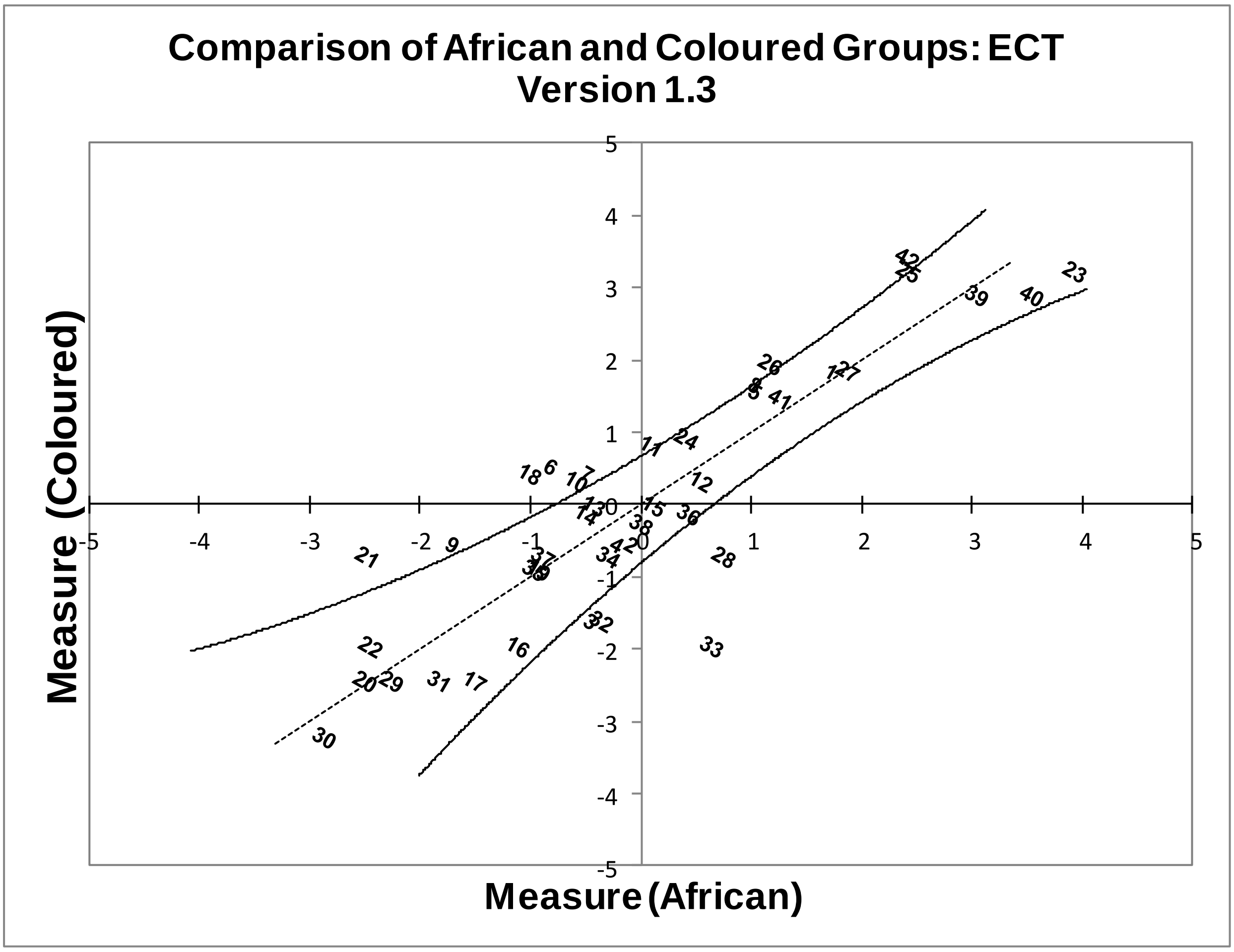


Figure 42: DTF Scatterplot for the African and Coloured Groups with Identity Line for the ECT version 1.3

Table 63: DTF Statistics for the African and Coloured Group Comparisons for the ECT version 1.3

ECT Items	African Group Measure	African Group S.E.	Coloured Group Measure	Coloured Group S.E.	*t* statistic
1	1.73	0.09	1.83	0.31	-0.30979
2	-0.08	0.08	-0.57	0.39	1.230783
3	-0.45	0.09	-1.63	0.55	2.117295
4	-0.21	0.09	-0.57	0.39	0.899438
5	1.04	0.08	1.55	0.31	-1.59297
6	-0.81	0.09	0.51	0.32	-3.97094
7	-0.48	0.09	0.41	0.32	-2.67737
8	1.04	0.08	1.64	0.31	-1.87409
9	-1.7	0.12	-0.57	0.39	-2.76931
10	-0.58	0.09	0.31	0.33	-2.60194
11	0.1	0.08	0.8	0.31	-2.18643
12	0.55	0.08	0.31	0.33	0.7068
13	-0.42	0.09	-0.03	0.35	-1.07918
14	-0.49	0.09	-0.15	0.36	-0.91625
15	0.12	0.08	-0.03	0.35	0.417797
16	-1.11	0.1	-1.98	0.63	1.363878
17	-1.5	0.11	-2.46	0.76	1.250131
18	-1	0.1	0.41	0.32	-4.20568
19	-0.92	0.1	-0.92	0.44	2.46E-16
20	-2.49	0.16	-2.46	0.76	-0.03863
21	-2.47	0.16	-0.74	0.41	-3.9308
22	-2.44	0.16	-1.98	0.63	-0.70769
23	3.94	0.2	3.21	0.42	1.569258
24	0.42	0.08	0.9	0.31	-1.49927
25	2.43	0.11	3.21	0.42	-1.79655
26	1.18	0.09	1.93	0.32	-2.25621
27	1.88	0.1	1.83	0.31	0.153501
28	0.76	0.08	-0.74	0.41	3.590819
29	-2.25	0.15	-2.46	0.76	0.271086
30	-2.86	0.19	-3.24	1.04	0.359436
31	-1.82	0.13	-2.46	0.76	0.83005
32	-0.35	0.09	-1.63	0.55	2.296726
33	0.65	0.08	-1.98	0.63	4.141347
34	-0.29	0.09	-0.74	0.41	1.072036
35	-0.96	0.1	-0.92	0.44	-0.08865
36	0.44	0.08	-0.15	0.36	1.599862
37	-0.88	0.1	-0.74	0.41	-0.33174
38	0.01	0.08	-0.29	0.37	0.792498
39	3.05	0.14	2.88	0.39	0.410264
40	3.55	0.17	2.88	0.39	1.574837
41	1.27	0.09	1.45	0.3	-0.5747

42	2.42	0.11	3.4	0.45	-2.11549

In the DTF statistics, Table 63, the items that have the highest *t* statistic values are items 3 (2.12), 6 (-3.98), 7 (2.68), 9 (-2.77), 10 (-2.60), 11 (-2.19), 18 (-4.21), 19 (2.46), 21 (-3.93), 26 (-2.26), 28 (3.60), 32 (2.30), 33 (4.14) and 42 (-2.12). These items all fall outside of the 1.96 value, are thus significantly different, and can be classified as DIF items. These items fell outside of the 95% confidence lines for both the empirical trend and identity trend line scatterplots.

Table 64: DTF Statistics for the African and Coloured Group Comparisons for the ECT version 1.3

Statistics	**African Group**	**Coloured Group**
Mean	0.000476	0.000476
S.D.	1.613967	1.727788
Identity trend	3.342232	3.342232
Identity trend	-3.34128	-3.34128
Empirical trend	3.228411	3.456053
Empirical trend	-3.22746	-3.4551
Identity intercept	0	
Identity slope	1	
Empirical intercept	3.13699E-05	-3.35821E-05
Empirical slope	1.070522489	
Correlation	0.873490244	
Reliability	0.995368231	0.926048742
Disattenuated correlation	0.909806945	

In Table 64, the empirical slope is 1.0705, which differs from the origin. This makes the empirical and identity slopes different in that the item locations are not the same across groups. The correlation between the African and Coloured groups is .873, which is not a satisfactory correlation due to the item locations that are not equivalent. The reliability value for the African group is excellent (.99), while the Coloured group also has a very good reliability (.926) (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009). The high reliabilities suggest that there is a good internal consistency present for these two groups. The disattenuated correlation of .909 is good and leaves .081 for a small extraneous variable that could be part of this relationship.

7.7 Reliability Analysis Results

To assist in the interpretation of the reliability analysis results, the output will be reported as follows: the reliability results and the revised reliability results. The results are not presented separately for the two test versions, as it is more valuable to simultaneously present these values for the two test versions.

The reliability analysis for the two versions was observed by the Kuder-Richardson Formula 20, which was calculated in SPSS 23, is also indicated below. The results for the two versions are presented in Table 65 below.

Table 65: Reliability Results for the ECT version 1.2 and 1.3

Test Version	**Kuder-Richardson Formula 20**	**Sample Size**	**Mean**
ECT version 1.2	*p* = .789	39	23
ECT version 1.3	*p* = .785	42	26

According to ranges for determining acceptable *p* values, the values for both test versions are within the .7 - .79 range (Table 65). They are, however, not reliable enough for aptitude tests or selection purposes. Since they are relatively close to .80 (which is the acceptable range for an aptitude test), there are possibly items within the test that may be lowering the reliability, and if removed, may produce a test reliable enough for aptitude testing purposes.

To assess the best reliability coefficient for the data, the item-total statistics were reviewed. These statistics highlight the items that either increase or decrease the reliability coefficient value. For the best coefficient to be obtained, the items that decrease the reliability coefficient were deleted and the reliability analysis was rerun. This process was repeated until the reliability coefficient was at its highest value (Table 66). For the ECT version 1.2 (Table 66), the following items were deleted to improve the *p* value: item 9.2, item 9.3, item 9.4, and item 9.5. For the 35 items remaining, Kuder-Richardson Formula 20 on standardised items was .812. This value means the test is reliable enough for aptitude testing (Nunnaly & Bernstein, 1994). It is, however, still insufficiently reliable for selection purposes or high-stakes testing (Foxcroft & Roodt, 2009). The mean for these 35 items was 21, which is slightly less than the entire sample mean of 23.

Table 66: Revised Reliability Results for the ECT version 1.2 and 1.3

ECT Test Version	**Reliability Statistic**	**Sample Size**	**Mean**
ECT Version 1.2	.812	35	20.86
ECT Version 1.3	.810	35	21.70

In the ECT version 1.3 (Table 66), the following items were deleted to improve the *p* value: item 6, item 7.2, item 7.3, item 7.4, item 7.5, item 14.1, item 14.6, and item 15.2. There are 35 items remaining and Kuder-Richardson Formula 20 on standardised items was .810. This value indicates that the test is reliable enough for aptitude testing (Nunnaly & Bernstein, 1994). It is, however, still insufficiently reliable for selection purposes or high-stakes testing (Foxcroft & Roodt, 2009). The mean for these 35 items was 22, which is a little less than the entire sample mean of 26.

7.8 Conclusion

The chapter focused on the analyses that were indicated in Chapter 1 and Chapter 6. These analyses were performed to achieve the objectives of the study, ultimately assisting in achieving the overall aims of the study.

Within this chapter, the results were presented for each test version for the different objectives, except for the ECT 1.2 in the CFA analyses. The reason for this was, however, explained. The presentation of results for both test versions allows one to explore the similarities and differences between the two versions when the same analyses were conducted.

The presentation of these results is an essential part of the study and provides an indication of the data and the information that can be obtained to establish the construct validity of the ECT. These results were informative and provided stimulating insights into the nature of the ECT as well as the construct of verbal reasoning being measured by the ECT. These results are thus imperative to the discussion, which follows in the next chapter.

CHAPTER 8: DISCUSSION OF RESULTS

8.1 Introduction

In this chapter, the results presented in the previous chapter are discussed to make reasonable conclusions about the construct validity of the ECT. The various analyses performed in the previous chapter were conducted to achieve the objectives of the study and thereby realise the study's aim, which was to evaluate the construct validity and reliability of the ECT. Since the results obtained from the analyses provided great insights into the test's development from the one version to the next, the discussion section requires three approaches to discuss the information obtained by these results.

The first section of the discussion will compare the results obtained from the analyses of the two test versions of the ECT. This is essential, as the researcher needs to ascertain whether the newer test version is indeed an improvement on the initial test version. This comparison is crucial to exploring the construct validity and reliability of the ECT and thus these results need to be evaluated effectively. Additionally, reference will be made to relevant literature to place the results in context.

The second section of the discussion will involve the use of Messick's unified theory of construct validity, with a specific focus on the six facets of construct validity. These six facets are the theoretical framework from which the results will be interpreted. This will allow for a greater understanding of the results obtained from the various analyses and will establish whether the ECT measures verbal reasoning, thereby providing evidence of the construct validity and reliability of the ECT.

The third section consequently contains an argument regarding verbal reasoning and will make reference to literature and the findings of the study as a whole. This is crucial to the exploration of the construct validity of the ECT.

8.2 Rasch Analyses Discussion

For a suitable comparison to be made, the performance of the items and persons on the two versions will be evaluated in terms of the separation values, infit and outfit MNSQ, empirical randomness graph, maximum and minimum measures, item-person map, measure order, misfit order, dimensionality, characteristic curves and information function results.

The person separation value for the ECT version 1.3 was slightly lower than the ECT version 1.2. These values are, however, both small and indicate that the persons completing the ECT did not have a good range of abilities. Thus, there were too many individuals with the same ability level and there was not enough variation in their ability. The item separation value was smaller for the ECT version 1.2 than the ECT version 1.3. Both of these values are considered large and indicate that the items of the ECT have a good range of difficulties. The variation in items, therefore, improved from the one version to the next. This is a positive finding and is what one hopes to discover when comparing test versions. Thus, there are items with different ability levels; easy, moderate and difficult.

The maximum MNSQ infit values for the two versions were similar, although the ECT version 1.3 value was slightly lower than the ECT version 1.2 value. This would indicate that the items performed generally as expected. The maximum MNSQ outfit values for both versions were similar, but the ECT version 1.3's value was slightly higher. This suggests that some items were misfitting. The minimum MNSQ infit values were similar for the two test versions and suggest that there was no difference in the over-fit observed across versions. The minimum MNSQ outfit value for the ECT version 1.2 was slightly higher than the ECT version 1.3. This means that the difference between the two test versions was very small in terms of the over-fit items.

When comparing the test empirical graph for the two versions, the outfit MNSQ was the greatest contributor to test randomness for both versions. This corresponds to the fit statistics observed, which suggested that the outfit MNSQ seems to be problematic. This can be interpreted as external factors which may have affected the test.

The maximum measure value for the ECT version 1.3 was much higher (3.77) than the ECT version 1.2 (2.78). This implies that the items in the ECT version 1.3 had a higher ability level than those in the ECT version 1.2. The minimum measure value for ECT version 1.2 (- 3.04) was much lower than the ECT version 1.3 (- 2.85). This means that there were items in the ECT version 1.2 that were measuring much lower ability levels than the items in the ECT version 1.3. The measure values are, however, more useful when compared to the person measure values. This comparison indicates the suitability of the items for the persons completing the different test versions.

When comparing the item-person map across test versions, the ECT version 1.2 and 1.3 had very similar results. There was an overall good fit and spread of items and persons, and the items had a better spread of ability than the persons. There were a few gaps in the continuum for both persons and items. There were items that had the same difficulty, which made them redundant. The majority of candidates fell between -1 and 2 standard deviations (average to above average). Thus, the persons of the test were mostly of the same ability level. This was observed for both test versions.

When observing the measure order statistics across the two test versions, there were some similarities in terms of the items identified. For the ECT version 1.2, the very difficult items were items 36, 37, and 39, while the easiest items were items 10, 12, and 13. For the ECT version 1.3, the very difficult items were items 23, 25, 39, 40, and 42, while the easiest items were items 20, 22, 29, and 30. Interestingly enough, there were no items that the candidates could not answer, and no items were correctly answered by all candidates. This was observed for both test versions.

Within the misfit order, there were some different misfit items across the two test versions. For the ECT version 1.2, the items that had MNSQ outfit values greater than 1.3 (Bond & Fox, 2007; Pensavalle & Solinas, 2013) were items 12, 16, 17, 18, and 19. Item 37 had an MNSQ outfit value smaller than 0.7 (Bond & Fox, 2007; Pensavalle & Solinas, 2013). There were no poor MNSQ infit values, and there were problematic correlations observed for items 16 and 17. For the ECT version 1.3, the items that had MNSQ outfit values greater than 1.3 (Bond & Fox, 2007; Pensavalle & Solinas, 2013) were items 6, 8, 9, and 23. The items that had MNSQ outfit values smaller than 0.7 (Bond & Fox, 2007; Pensavalle & Solinas, 2013) were items 30 and 40. There were no poor MNSQ infit values, and there were problematic correlations observed for items 8, 9, and 23.

The variance decomposition for the two test versions was different, but the amount of dimensions identified was the same. For the ECT version 1.2, the raw unexplained variance was 39 (68%). The variance explained by the measures was 18 (32%), and the four dimensions identified were 1 (3.08, 5%), 2 (2.10, 4%), 3 (1.58, 3%), and 4 (1.42, 3%). For the ECT version 1.3, the raw unexplained variance was 42 (64%). The variance explained by the measures was 23 (36%), and the four dimensions identified were 1 (2.90, 4%), 2 (2.13, 3%), 3 (1.68, 3%), and 4 (1.52, 2%).

The standardised residual plots for the two test versions differed in terms of which items lay where, but the plot shape looked very similar across the two test versions. For the ECT version 1.2, items A (item 29), B (item 31), C (item 28), D (item 32), and E (item 30) were identified in the standardised residual plot as different from the other items in the test. For the ECT version 1.3, items A (item 37); B (item 35), C (item 34), D (item 38) and E (item 36) were identified in the standardised residual plot as being separate from the other items in the test. This finding contradicts with the EFA findings, because in the EFA, these items were related (Arendse & Maree, 2017) and were identified as factor 1 in the principal axis factoring analysis. This factor was labelled Vocabulary, due to the nature of the items (antonyms). Although the items differ across the two test versions in the standardised residual plots, the sets of items identified corresponded with the respective EFA for each test version.

The standardised residual loadings were similar across the two test versions, in terms of the item content for items A, B and C (antonyms). For the ECT version 1.2, the top three items identified were A (item 29), B (item 31), and C (item 28). The bottom three items identified were a (1), b (23), and c (15). For the ECT version 1.3, the top three items identified were A (item 37), B (item 35), and C (item 34). The bottom three items identified were a (13), b (26), and c (27).

The test characteristic curve for the two test versions had an s-shape, which indicated that there was a fair fit to the model. The steepness of the curve indicated that there was a relative range of difficulty. Redundancy was present, as well as over-fit items that were not yielding new information. The test does; however, seem to be discriminating between low and high performers. This was observed for both test versions.

The item characteristic curves for the two test versions were different in that the items observed differed according to their difficulty level in the respective test version. For the ECT version 1.2, the easy item was item 10, the moderate item was item 5, and the difficult item was item 36. For the ECT version 1.3, the easy item was item 30, the moderate item was item 12, and the difficult item was item 39.

The test information function for the two test versions were similar in that curve peaked at 7.5 logits and followed a range of -7.6 to 7.6 logits. This implied that there was a fair probability of easy and difficult items. The width of the curve indicated that there was a relative measurement range present in the test.

In light of these critical findings of Rasch analyses for the two test versions, aspects regarding the sample connect to the literature explored. Since the persons were problematic for both test versions and it was found that they were misfitting and causing unexpected response styles, reasons for this can be attributed to several possible explanations. Firstly, words may have social references (Radden, 2008) and this may cause unexpected responses for persons. Secondly, changes in the education system (Koch, 2015) may have impacted their performance on the items. Thirdly, Piaget’s initial interest in exploring incorrect responses (Santrock, 2010) could be used to explain the same issue (incorrect and unexpected responses by persons) as he attributed this to either boredom (items are too easy) or withdrawing (items are too difficult) (Santrock, 2010). Fourthly, according to Piaget’s formal operational stage of cognitive development, persons may inconsistently use formal operational thought and may struggle with abstract reasoning if they have not successfully progressed through this stage (Blake & Pope, 2008; Wankat & Oreovicz, 1993). This may explain the observed incorrect and unexpected responses to items. Fifthly, the consideration of time is important in intelligence measures (Keith & Reynolds, 2010). Since the ECT can be considered to be measuring a form of intelligence, the improvement observed for the ECT version 1.3 could be due to the removal of the time limit.

8.3 Differential Test Functioning Analysis Discussion

8.3.1 Discussion of Gender

The similarity across genders, therefore, suggests that the samples were problematic in terms of their variation of abilities across test versions. The items, however, improved in terms of the variation of abilities being tested. They further showed promising results across genders, concerning high reliability and good average infit and outfit MNSQ values.

There were five DIF items identified for the ECT version 1.2, namely items 4, 20, 24, 33, and 39. There were six DIF items identified for the ECT version 1.3, namely items 4, 6, 26, 27, 32, and 36. The common DIF item identified for both ECT versions was item 4 (true or false item).

The findings on gender for the ECT version 1.2 (only the ECT version 1.2 had a time limit) indicated that the males performed better than females. This contradicts the findings of Griskevica & Rascevska (2009) as the females did not outperform the males with a time limit on verbal assessment (Griskevica & Rascevska, 2009). It should, however, be cautioned that there were much fewer females (27%) than males (66%) in the ECT (version 1.2) sample and thus this deduction drawn from the findings would not be a fair one in terms of the gender distribution.

8.3.2 Discussion of Racial Groups (African, White and Coloured)

There was a similarity in terms of the limited variation of person abilities across the three racial groups which suggested that the candidates were problematic across test versions. Additionally, the person reliability across the race groups was not good and could have been higher. The items, however, improved in terms of the range of difficulty being tested across the three race groups. They furthermore showed promising results across the three race groups, with regards to high reliability and good average infit and outfit MNSQ values.

There were 13 DTF items identified for the ECT version 1.2 between the African and White group, namely items 2, 6, 9, 12, 14, 17, 18, 19, 24, 27, 36, 37, and 39. There were 18 DTF items identified for the ECT version 1.3 between the African and White group, namely items 2, 6, 7, 8, 9, 10, 11, 13, 18, 21, 24, 25, 26, 28, 32, 33, 39, and 40. There were nine common DTF items identified for the ECT version 1.2 and ECT version 1.3. These were items 2, 6, 12, 17, 18, 19, 24, 36, and 37 for the ECT version 1.2 and items 2, 6, 9, 10, 11, 21, 28, 39, and 40 for the ECT version 1.3.

There were nine DTF items identified for the ECT version 1.2 between the African and Coloured group, namely items 6, 8, 9, 12, 14, 17, 18, 19, and 24. There were 14 DTF items identified for the ECT version 1.3 between the African and Coloured group, namely items 3, 6, 7, 9, 10, 11, 18, 19, 21, 26, 28, 32, 33, and 42. There were seven common DTF items identified for the ECT version 1.2 and ECT version 1.3. These were items 6, 8, 12, 17, 18, 19, and 24 for the ECT version 1.2 and items 6, 9, 10, 11, 18, 21, and 28 for the ECT version 1.3.

Based on these DTF statistics, it was necessary to assess which DTF statistics were common across each of the test versions to locate the problematic items. The common DTF items for the ECT version 1.2 across genders and the three race groups were items 6, 9, 12, 14, 17, 18, 18, 24, and 39. The common DTF items for the ECT version 1.3 across genders and the three race groups were items 6, 7, 9, 10, 11, 18, 21, 26, 28, 32, and 33.

A way in which the differential test functioning (DTF) can be resolved is to disregard the DTF as intrinsic to the measurement system of the test. In the ECT, the majority of the test takers were not English first language speakers, and for this reason, different items will exhibit DTF for the speakers of different languages. This DTF is also dependent on the relationship of the test takers native languages to the English language (Linacre, 2015). In the DTF analyses, the comparison was based on racial groups; primarily the majority racial groups (African, White, and Coloured). Furthermore, certain languages are associated with African members, as there are nine official languages (Xhosa, Ndebele, Tsonga, Zulu, Siswati, Setswana, Sepedi, Sotho, and Venda) within this group. The White and Coloured group commonly consists of a combination of the remaining two official languages (English and Afrikaans). This would, therefore, explain the differences observed in the analysis of the DTF for the racial groups. The relationship between these racial groups and the possible bias exhibited for each of these comparisons may be due to the differences in their languages when compared to English. The Afrikaans language may be more similar to the English language than the different African languages. Thus the differences observed across the race groups are due to the differences in languages (African languages and Afrikaans) when compared to the English language and not specifically racial differences. Additionally, it should be noted that different racial groups tend to perform differently on vocabulary assessments. The reasoning behind this difference in performance is linked to similar studies with African Americans where differences in vocabulary were attributed to variances in culture, language, and social experience (Pae et al., 2012). These explanations may serve to contextualise the findings of the DTF and may provide a more comprehensive outlook on the meaning of the results and their consequent implications. Socio-economic contexts are often the basis of differential performance of children on cognitive assessments, specifically in South Africa (Cockcroft et al., 2016).

The findings observed for the different gender and racial groups across the two test versions contain elements that link to the discussed literature. These aspects may serve to explain the differences observed between these groups, such as the social meaning attached to certain words (Radden, 2008) and how individuals may differ in their thinking due to language differences and semantic structures (Boroditsky, 2011; Gentner & Goldin-Meadow, 2003). These differences may also connect to Vygotsky's emphasis on the influence of culture and language on cognitive development (Ormrod, 2008), as the environment and learning opportunities can influence this (Van der Pool & Catano, 2008).

8.4 Multi-Trait Multi-Method Discussion

There were several common themes that emerged when observing the correlations of the ECT version 1.2 and 1.3, despite the differences between the two test versions. This is very important as it suggests that an inherent construct is present in the ECT. The psychometric tests (constructs) that were theorised to be more closely related to the ECT were indeed related and shared the strongest relationship. These prominent constructs were verbal reasoning, vocabulary, reading comprehension, and long-term memory. Literature also indicated that these constructs are closely related due to the reasoning element common among them. These constructs had the strongest relationship to each other as well as to the ECT. This provides evidence of the theorised existence of verbal reasoning as the core construct of the ECT, and within this construct several other related constructs, such as vocabulary, are present. These correlations also indicated, to some extent, the existence of convergent validity, as the strongest relationships observed among the constructs were the hypothesised relationships.

The psychometric tests (constructs) that were hypothesised to be less or not related to the ECT were the following: Non-Verbal Reasoning, Numeric Comprehension, Mechanical Insight, Calculations, Comparison, Pattern Completion, Figure Series, Spatial 2D, and Spatial 3D. The correlations of these psychometric tests (constructs) suggested that there was a relationship observed between the ECT and these constructs. These correlations were, however much smaller than those observed between the hypothesised constructs. A possible reason for the existence of a relationship between these constructs and the ECT is that reasoning is common to all these constructs (Marshalek, 1981). This makes it difficult to confidently confirm convergent and discriminant validity, as there are relationships observed across all the constructs for both test versions. The reason for this is that these forms of reasoning are interrelated. If one only focuses on the verbal reasoning correlations, they range from moderate to large and may partially indicate convergent validity. The reason for this is that the strongest relationships observed are between the constructs hypothesised to be related. To confirm discriminant validity, one needs to observe low correlations with constructs that are less related to the construct being measured. The fact that the relationship was smaller between the ECT and the constructs hypothesised to be less related allows one to argue that discriminant validity is partially obtained.

The time limit imposed on the test versions could have contributed to the difference in correlations noticed with the memory tests. This difference could be attributed to a difference in memory retrieval when a time limit was imposed. These correlations emphasised the differences in memory required for the tasks across the test versions of the ECT. This is based on the link between memory and analogical retrieval (Holyoak, 2012). Working memory is associated with fluid intelligence and is related to language comprehension and reading. Research conducted in South Africa with children concerning verbal working memory and vocabulary indicated that verbal working memory was less prejudiced towards the socio-economic status of individuals. Vocabulary, however, was heavily influenced by the socio-economic status of individuals. This emphasises the fact that educational opportunities that are linked to socio-economic status have an impact on how individuals perform on vocabulary assessments (Cockcroft, Bloch, & Moolla, 2016). Moreover, the relationship between vocabulary and different forms of reasoning observed in the results may relate to crystallized intelligence (Cockcroft et al., 2016; Marshalek, 1981).

The correlations observed between all the psychometric tests and the ECT may suggest the following: Verbal ability links to fluid and crystallized intelligence (Kvist & Gustafsson, 2007); verbal ability (the verbal assessments and the ECT) may be a good indicator of general intelligence due to its relation to all the tests (Kvist & Gustafsson, 2007); assessing intelligence is unavoidable when measuring ability (Pelser, 2009); and these relationships across the various tests suggest that intelligence could be viewed as a method of processing (Fagan, 2000) or mental speed and working memory (Almeida et al., 2011; Sternberg, et al., 2011; Taylor, 1994). Moreover, these findings may be related to connectionism, which relates cognition and language (Harris, 2006).

8.5 Confirmatory Factor Analysis Discussion

In the CFA analysis of the ECT version 1.3, there were several important findings such as the reasoning and education factors that had the largest impact on verbal reasoning, because they explained the largest amounts of variance for verbal reasoning (indicated by the standardised regression weights). The factor correlations indicated that the education and reasoning factors accounted for the largest variance in the model. The other factors: deduction, plurals and vocabulary accounted for smaller variances in the model. These factors are consequently necessary, as they form part of the construct underlying the ECT. The model fit statistics (chi-square, TLI, CFI, RMSEA, RMR, AIC and CAIC) all indicated that the hypothesised model was acceptable and that the data was a good fit to the model.

The CFA structure (graphical input) created for the ECT version 1.3 which was mostly based on the EFA conducted on the ECT version 1.2 was confirmed and therefore the structure can be regarded as consistent across test versions. Thus, regardless of time and test structure differences between the two test versions, the ECT has a particular set of dimensions present. This is emphasised by the good model fit indices. Overall, the CFA presented positive results concerning the dimensionality of the ECT as it validated the hypothesis that the structures of the two test versions would be similar.

The education factor, which was created for the CFA analysis of the ECT version 1.3 from the items that did not load on any of the other factors in the EFA of the ECT version 1.2, can be regarded as crystallized intelligence. The reasoning factor, which was identified in the EFA of the ECT version 1.2 and was a factor in the CFA of the ECT version 1.3, can be considered fluid intelligence (Holyoak, 2012). Crystallized and fluid intelligence form a critical part of the verbal reasoning construct. These two forms of intelligence (which can be regarded as the education and reasoning factors) were identified as having a strong relation to the ECT (verbal reasoning) as a verbal assessment (Kvist & Gustafsson, 2007). Additionally, the positive relationship observed between vocabulary and reasoning confirms findings in literature (Marshalek, 1981).

The identification of the education factor as the strongest factor (explaining the most variance) of the verbal reasoning construct (which may be viewed as a form of intelligence) may suggest a connection to the investment hypothesis of intelligence by Cattell. This hypothesis proposes that cognitive abilities are influenced by the environment and learning (Van der Pool & Catano, 2008). This hypothesis supports the relevance of the education factor as a facet which contributes to the construct of verbal reasoning.

8.6 Reliability

The reliability coefficient of .78 was observed for both test versions, regardless of the changes across these test versions and differences in administration of the ECT. Although the ECT version 1.3 has a higher mean and more items, the reliability coefficient was the same as the ECT version 1.2. The higher mean implies that candidates performed better on the ECT version 1.3 than on the ECT version 1.2, which could be because there was no time limit imposed on the ECT version 1.3. The Kuder-Richardson Formula 20 of .78 is a good reliability value, but has restrictions in terms of its use for making selection decisions and sufficiently measuring aptitude (Nunnaly & Bernstein, 1994). It does, however, suggest that there is relatively good internal consistency across the two test versions, as both test versions, regardless of differences, yielded similar results.

The reliability coefficient improved for both test versions when certain items, based on their item-total correlations, were removed. This improvement was similar across both test versions, as the amount of items remaining (35) and Kuder-Richardson Formula 20 (.81) remained the same. The only difference between these test versions after the improvement of Kuder-Richardson Formula 20 was that the mean was still higher for the ECT version 1.3. The Kuder-Richardson Formula 20 of .81 suggests that this test is able to sufficiently assess aptitude, yet falls short of being appropriate for selection decisions, which are high-stakes (Nunnaly & Bernstein, 1994). This improvement is good and suggests that the ECT adequately measures aptitude, which allows the test to be useful for aptitude-related decisions. The test can, however, not be used for high-stakes decisions such as selections. The fact that this improvement was observed across both test versions suggests that this improvement is consistent and therefore the internal consistency of the test is reliable.

When reviewing these reliability coefficients across test versions, the Kuder-Richardson Formula 20 for both test versions was consistent even when improved by deleting items. Additionally, the good reliability is a prerequisite for unidimensionality, and thus it would appear that the ECT could be measuring verbal reasoning consistently. This suggests that there is little measurement error in the assessment of the ECT. This also suggests that the ECT can assess the construct of verbal reasoning fairly accurately. The measurement error is also low enough for aptitude assessment, since verbal reasoning is considered an aptitude.

8.7 Messick’s Theoretical Framework

This study has utilised Messick’s theory of construct validity to integrate the various results obtained. This integration allows one to argue whether the ECT has sufficiently displayed evidence of construct validity, especially when assessing the six facets by which this theoretical framework is guided. These six facets are called content, substantive, structural, generalisability, external, and consequential aspects (Messick, 1995, 1996; Ravand & Firoozi, 2016; Smith, 2001). All the analyses conducted do not relate to each of the facets and for this reason, only the analyses that was able to add value to the facet of validity being discussed will be identified.

The analysis techniques that addressed these facets of construct validity are identified as follows: Rasch Analyses (content, substantive, structural, and generalisability), DTF (content, generalisability, and consequential), reliability (content, structural, and generalisability), CFA (content, structural, and generalisability) and MTMM (generalisability and external).

8.7.1 Content Facet of Construct Validity

The content facet of construct validity addresses the content of the test and assesses whether the content is appropriate for construct being measured (Baghaei & Amrahi, 2011; Messick, 1995, 1996; Ravand & Firoozi, 2016; Smith, 2001). For the ECT, the content facet would establish whether the construct of verbal reasoning is sufficiently assessed by the content in the ECT. The analyses that assisted in assessing the content facet of construct validity were Rasch analyses, DTF, CFA, and reliability (Baghaei, 2008; Baghaei & Amrahi, 2011; Ravand & Firoozi, 2016; Smith, 2001).

The Rasch analysis was able to adequately integrate Messick's theory of validity because of the variety of outputs it provided. The content facet of construct validity was addressed by examining the fit indices of the Rasch analyses. The reasoning behind this was that it assures one that the items used in the test were acceptable and were connected to the construct that was measured. The identification of misfit items indicated that there was construct-irrelevant information being measured in the test (Baghaei, 2008; Baghaei & Amrahi, 2011; Ravand & Firoozi, 2016; Smith, 2001).

In the ECT, the fit indices for the items were determined for both versions. The average infit and outfit MNSQ and ZSTD values were acceptable across both test versions and indicated that, on average, the items fit the model. The maximum and minimum infit and outfit MNSQ and ZSTD values across both test versions indicated that there were misfitting and over-fitting items in the ECT. These items were indicative of problematic fits to the model, and therefore do not measure content-relevant information. The item separation values across both test versions showed an improvement in terms of the range of difficulty measured by the items, which indicated that the items sufficiently addressed the content in terms of the different difficulty levels. The standard error of measurement was small across both test versions, indicating that there was little error observed in the measurement of the items of the ECT. The item reliability was very high and indicated that the items measured content-relevant information (Baghaei & Amrahi, 2011; Ravand & Firoozi, 2016).

The item-person map was also essential in establishing whether the content used in the items was relevant. The representivity of the items and persons was visible in the map. The identification of gaps should be regarded as areas in which the construct has not been adequately measured (Baghaei, 2008; Baghaei & Amrahi, 2011; Ravand & Firoozi, 2016; Smith, 2001). The item-person map had very similar results across test versions. There was an overall good fit and spread of items and persons, with the items having a better spread of ability than the persons. There were a few gaps in the continuum for both persons and items and some items had the same difficulty, making these items redundant. The item-person map therefore indicated that the content was sufficiently measured by the ECT, as most of the items were well spread.

The technical quality of the content can be addressed by assessing the fit indices and item measure correlations. The fit indices allow one to establish whether there is multidimensionality or unidimensionality present and if questionable item quality is observed (Baghaei, 2008; Baghaei & Amrahi, 2011; Ravand & Firoozi, 2016; Smith, 2001). The item measure correlation allowed one to determine that the items had a relation to the test as a whole. These correlations should be positive so that the relationship between the items is positive. Low correlations (those close to zero) suggest that the item is possibly very difficult or very easy, and is therefore not easily endorsed, which may mean that the item does not measure the construct as well as the other items in the test (Baghaei & Amrahi, 2011; Ravand & Firoozi, 2016). A similar pattern was observed in the ECT across test versions; mostly low to moderate correlations were observed. Some items were more related to the construct than others. This finding relates to what was observed by the fit indices as well, as most items were related to the construct.

The results of the various Rasch analyses such as the fit indices, item-person map, and the item measure correlations suggested that most of the items did not cause concern and appeared to be measuring the content facet of construct validity. There were some items that have been noted across the two test versions that posed a threat to the content of the ECT, but overall, the content appeared to be measured effectively.

The DTF analyses contributed to the understanding of the content facet of construct validity. The information obtained by these analyses across the test versions was vital as it assisted in establishing whether different races and genders experienced the content similarly. In the DTF analyses for males and females, the fit statistics were observed to assess whether the two samples had similar or varied abilities. The average infit and outfit MNSQ for persons and items were acceptable across test versions. Based on the person separation values, it was observed that there were similarities across genders for both test versions, which suggested that the persons completing the ECT were problematic in terms of their variation of abilities. The item separation values, however, demonstrated an improvement regarding the variation of difficulties being tested. The values further showed promising results across genders with regards to high reliability.

The DTF produced scatterplots, which assisted in establishing whether the two genders responded similarly to the various items. This similarity in response was observed in the scatterplot, as there were only a few outliers. These outliers were then assessed by the *t* statistics to establish whether they were possibly DIF items. A similar trend was observed for both test versions: The genders seemed to respond to items similarly and only a few items had high *t* statistic values, indicating possible DIF present. Based on how the two genders performed on the ECT for both test versions, one could argue that the content was understood in a similar way and most items did not discriminate against either gender. This suggests that the test content for both test versions was appropriate for both genders, excluding those items identified as possibly biased (Bond & Fox, 2007; Linacre, 2012d).

The DTF conducted on the three racial categories (African, White, and Coloured) indicated that there were some similarities observed in the fit indices across the groups. The average infit and outfit MNSQ for both persons and items were acceptable across test versions. These person separation values suggested that the persons of these three racial samples were problematic in terms of their variation of abilities across test versions. The reliability value could also have been higher. The item separation values for the three racial categories, however, demonstrated an improvement regarding the variation of difficulties being tested across the three racial groups. The item separation values further showed favourable results across the three race groups and had a high reliability value. The DTF produced scatterplots for the different races which assisted in establishing whether the three groups responded similarly to the different items. This was observed in the scatterplot by comparing the White and Coloured groups to the African group, as this was the dominant group in the sample. There were a number of outliers observed on these scatterplots. These outliers were assessed by evaluating the respective *t* statistics to establish whether they were possibly DIF items.

There was a similar trend observed for both test versions; the racial groups performed similarly across test versions. The White and Coloured groups performed differently compared to the African groups. These racial group differences seem to have been consistent over the test versions, with some items appearing biased. A possible reason for these biased results could be attributed to the wording of those items, but this would require further investigation. Based on how the White and Coloured groups performed when compared to the African group, it becomes apparent that there are a few more items that could be biased towards the race groups. These items, however, need to be explored in DIF to investigate the bias more thoroughly.

Overall, the Coloured and White groups performed similarly to the African group on the majority of the items of the ECT for both test versions. There were, however, items of each version that need to be inspected to ascertain why the groups performed very differently on these items. This differential performance could be due to language differences, specifically the differences in how different language speakers reason in their home language. As most of the individuals speak English as a second language, it is expected that some items would be biased. The fact that all three racial groups performed similarly on most of the items for both test versions suggests that the content was understood in a similar way and most items were not discriminating against their racial grouping. This would therefore exclude the performance on items that exhibited high *t* statistics and are possibly biased. Additionally, these DTF results for the three racial groupings suggest that the test content of the ECT for both test versions was applicable, excluding those items identified as possibly biased.

The CFA results were also instrumental in establishing whether the content was appropriate for the construct of verbal reasoning being measured. The CFA analyses were based on the EFA analyses conducted on the ECT version 1.2 and were performed on the ECT version 1.3. The model created for the ECT version 1.3 was, therefore, based on the EFA model of the ECT version 1.2. The model fit statistics (chi-square, CFI, TLI, RMSEA, RMR, AIC, and CAIC) all indicated that the hypothesised model was acceptable and it fitted the data. This suggests two points. Firstly, there is a definite structure present in the ECT, regardless of test versions. Secondly, the differences in the test versions differ only by the addition of items and possibly the influence of time pressure, as this may have impacted performance on the items.

Furthermore, the reasoning and education factors accounted for most of the variance in the model, which was indicated by the standardised regression weights. This is important as it suggests that the content of the ECT is predominately influenced by these factors. The factor correlations indicated that the different factors, reasoning, plurals, vocabulary, and deduction, are separate dimensions present in the test. These factors are therefore necessary, as they form part of the construct underlying the ECT – verbal reasoning. Since the model fit and factor structure results are positive, one can argue that these results suggest a good fit between the content of the ECT and construct of verbal reasoning (Kline, 2011).

In the reliability analyses, the Kuder-Richardson Formula 20 results for both test versions were vital in establishing that the construct of verbal reasoning was measured by the content of the ECT. The Kuder-Richardson Formula 20 was consistent across both test versions. The *p* value of 0.78 suggested that most of the items were reliable and there was very little variance left which could be considered unexplained and affected by error (0.22). The revised Kuder-Richardson Formula 20 for both test versions was 0.81, as problematic items were removed. This value is considered a good reliability value and is suitable for measuring aptitude in assessments (Erguven, 2014; Nunnaly & Bernstein, 1994; Suhr & Shay, 2009). Moreover, a good reliability is a prerequisite for unidimensionality and thus it would appear that the ECT could be measuring verbal reasoning consistently.

These reliability results suggest that there is little measurement error contained in the assessment of the ECT. This may then imply that when assessing verbal reasoning, the ECT can assess this construct fairly accurate. In addition to this, the measurement error is low enough for aptitude assessment. This good reliability value, in conjunction with the finding that the content is being measured accurately, is a promising result.

The various analyses that were explored to establish whether the content of the ECT was sufficiently represented yielded promising results. The results across test versions signify that there is definite consistency in the content of the ECT. Based on the Rasch, DTF, CFA, and reliability results, there is a consistency in the identification of the content by most items, which indicates that the content facet of construct validity has been met and has been adequately displayed by these statistical techniques.

8.7.2 Substantive Facet of Construct Validity

The substantive facet of validity refers to the confirmation that the construct being measured in the test is, in fact, being assessed (Messick, 1995, 1996; Ravand & Firoozi, 2016; Smith, 2001). The way in which one assesses this is to establish that the persons completing the assessment have engaged with the specific construct. The substantive facet of validity can be addressed by the person-fit statistics. This facet is established by examining the pattern of the persons in correspondence to the model fit. If persons do not fit the model, it could be due to factors outside of the construct being measured. This would include individuals cheating, guessing, or being careless in their responses (Baghaei & Amrahi, 2011; Ravand & Firoozi, 2016; Smith, 2001).

When reviewing the results for the persons of the Rasch analyses, several sets of information must be considered to establish that the substantive facet has been realised. The statistics relating to the persons fit to the model involve the following: person separation, person reliability, person infit and outfit MNSQ, person infit and outfit ZSTD, maximum and minimum MNSQ, and ZSTD values (Ravand & Firoozi, 2016).

The results of the person separation for both test versions were relatively small and indicated that the persons completing the ECT did not have good range of abilities. This means that for both test versions, there were too many individuals with the same ability level and there was not enough variation in their ability. The person reliability for both test versions was in the 0.7 range, which means that there was still some error in the measurement of the person’s ability and performance. The person reliability also suggests that there was some consistency in how the persons performed on the items. The average infit and outfit MNSQ and ZSTD values were acceptable. This indicated that on average, persons performed acceptably on the items of the ECT in terms of their abilities.

The maximum infit MNSQ values of both test versions indicated that the persons were slightly more misfitting in the ECT version 1.3 than the ECT version 1.2. Both test versions, however, had misfitting persons. This is an aspect that needs to be explored further and recommendations based on this will be made for future studies. The maximum outfit MNSQ values for both test versions were similar, which suggests that this might be a consistent issue. This could, however, be influenced by the fact that the many of the persons had the same ability level. The minimum MNSQ infit and outfit values for the ECT indicated that there were more over-fitting items in the ECT version 1.3 than in the ECT version 1.2. There were, however, over-fitting items in both test versions. The maximum ZSTD values for both test versions indicated that there were significant differences across the person’s ability. The minimum ZSTD values for both test versions were fairly similar and indicated that there were no differences observed across test versions.

The maximum measure value for the ECT version 1.2 was higher than the ECT version 1.3, which implies that there were persons with higher abilities completing the ECT version 1.2 than those completing the ECT version 1.3. The minimum measure value for ECT version 1.3 was significantly lower than that of the ECT version 1.2, which means that the persons completing the ECT version 1.3 had lower ability levels than those completing the ECT version 1.2. When comparing these measure values to the items, it is apparent that the items of the ECT for both test versions were below the lowest ability individual. In terms of the highest ability level, the items were higher than the person's ability in the ECT version 1.3, while the person's ability was higher than the items in the ECT version 1.2.

These various elements associated with the persons of the ECT are important as they allow one to establish the degree to which the person-fit results are acceptable or require attention. Based on the results of the persons, there were persons who were over-fitting and misfitting the model. This means that the persons completing the ECT did not behave according to their ability level and thus answered in an irregular or contrasting method. The reasons for this could be that these persons did not concentrate when they were completing the items, or they cheated, or guessed. The person fit was, however, largely in acceptable ranges for both test versions. This satisfactory average performance of persons suggests that to some degree there is substantive validity present. The other results, however, suggest that there are serious issues within the persons that caused them to not fit the model. These factors therefore pose a threat to validity.

8.7.3 Structural Facet of Construct Validity

The structural facet of construct validity is dependent on how the test is scored, as this allows one to infer the constructs on the test. When tests are multidimensional in nature, they need to be scored in a way that assesses these various constructs. This often takes the form of multiple test scores. Unidimensional tests, however, rely on one test score, as there is only one construct being assessed (Baghaei & Amrahi, 2011; Messick, 1995, 1996; Ravand & Firoozi, 2016; Smith, 2001). The analyses that can assist in determining the structural facet of construct validity are Rasch analyses, CFA, and reliability.

The structural facet of validity can be addressed by assessing the fit statistics and dimensionality of the Rasch Analyses (Ravand & Firoozi, 2016). Fit statistics are assessed by determining whether the items are unidimensional and thus a total score may be used to quantify the test and the construct being measured. When assessing the results of the fit statistics, both test versions need to be considered to ensure that the same information was assessed. The item separation indicated that the items had a good range of difficulties and increased from the one test version to the next. This indicates that the items were addressing different difficulty levels. This also suggests that there were fewer gaps in the item-person continuum as more ability levels were covered in the ECT version 1.3. The item reliability was very good across both test versions and indicated that there was very little measurement error observed in the items. There was furthermore high internal consistency among the items.

The average MNSQ and ZSTD infit and outfit values indicated that the items were performing acceptably across test versions. The maximum MNSQ infit values were acceptable, but the maximum MNSQ outfit values indicated that there were items that were not performing as they ought to and were misfitting. This was observed across the two test versions. The minimum MNSQ infit values across both test versions were considered acceptable. The minimum MNSQ outfit values across test versions indicated that there were over-fitting items. The maximum and minimum ZSTD infit and outfit values indicated that there were significant differences observed with the items across both test versions. These item statistics indicated that there was general consistency observed with the items across both test versions. One can argue that this implies that there is possibly one dimension being assessed by the items, with the exception of some of the items not performing as expected. This implication contributes to the argument that the items suggest unidimensionality, which supports the structural facet of the construct validity.

Rasch analyses also allowed one to assess the dimensionality of the items and this was done by evaluating the results of the variance decomposition (Ravand & Firoozi, 2016). The results of the variance decomposition were different across the two test versions and this complicates the argument. The raw unexplained variance for the ECT version 1.2 was 39 (68%), while the ECT version 1.3 was 42 (64%). The variance explained by the measures for the ECT version 1.2 was 18 (32%) and for the ECT version 1.3, it was 23 (36%). These differences, however, indicated that the variance explained improved from the one test version to the next. The unexplained variance was nonetheless concerning as it was more than the explained variance (which was observed across the two test versions). Both test versions identified the four dimensions, which were 1 (3.08, 5%), 2 (2.10, 4%), 3 (1.58, 3%), and 4 (1.42, 3%) for the ECT version 1.2 and 1 (2.90, 4%), 2 (2.13, 3%), 3 (1.68, 3%), and 4 (1.52, 2%) for the ECT version 1.3. These dimensions suggest some similarity in the structure present from the explained variance across the two test versions.

The results of the dimensionality of the ECT were promising as the ECT version 1.3 was an improvement from the ECT version 1.2. The identification of multiple dimensions suggests multidimensionality, not unidimensionality. This can, however, be explained as verbal reasoning is a multifaceted construct and it is expected that several factors load under this. For this reason, the ECT can still be considered unidimensional as the underlying construct of verbal reasoning is expected to have factors loading within it. When combining the results of the dimensionality and the item statistics, one can argue that these results suggest that the structural facet of construct validity has been established. The use of a total score for the ECT is therefore justifiable as the underlying construct is unidimensional. Although there are certain concerns regarding the variance not explained by the measures and some misfitting and over-fitting items, the observed results suggest that the structural facet has been identified.

The dimensions of the ECT version 1.3 were explored in the CFA analysis, as the model of the ECT version 1.2 was the theoretical baseline. In addition to this, the CFA was explored by conducting a structural equation modelling program, as the structure of the ECT needed to be assessed. The results regarding the factors indicated that the reasoning and education factors accounted for most of the variance in the model, while the remaining factors, deduction, plurals, and vocabulary, accounted for less variance. This implies that these factors are all relevant when explaining the variance of the ECT version 1.3, but there are factors that related more strongly to the construct being measured. This would indicate that these factors are necessary, as they form part of the construct underlying the ECT.

The chi-square, TLI, CFI, RMSEA, RMR, AIC and CAIC all indicated that the hypothesised model was acceptable and there was a good fit to the model. These results therefore suggest that the structure found in both versions is similar and indicates the relevance of verbal reasoning as an underlying construct to the factors emerging from the structure. Based on these results of the ECT structure, the argument that the ECT has established the structural facet of construct validity is justified.

The reliability of the ECT, which was calculated using Kuder-Richardson Formula 20, suggested that the ECT was sufficiently reliable for research purposes and when problematic items were removed, it was reliable for aptitude purposes. This good reliability coefficient suggests that there is consistency within the items and one can argue that this reliability combined with the Rasch and CFA results suggests that the ECT is unidimensional. This, therefore, means that the ECT is possibly measuring verbal reasoning.

The different analyses that were explored indicated that the structural aspect of the ECT was sufficiently observed. Based on the Rasch, CFA, and reliability results, there is a clear indication that the structural facet of construct validity was established. This was determined by exploring the various aspects related to the structure and identifying the unidimensional nature of the underlying construct of verbal reasoning in the ECT across test versions.

8.7.4 Generalisability Facet of Construct Validity

The generalisability facet of construct validity allows one to confidently use the score obtained from the test and trust the implications and the deductions made from the test (Messick, 1995, 1996; Ravand & Firoozi, 2016; Smith, 2001). The generalisability facet can be established by assessing the invariance of items. This invariance would limit the generalisability of the construct. One, therefore, seeks to be able to speak broadly about the construct being measured and not merely the components of the construct identified in the test. This includes the invariance of persons for particular items in the test. This would typically be done by conducting a DIF analysis on problematic items (Baghaei & Amrahi, 2011; Ravand & Firoozi, 2016). The generalisability facet of the construct will therefore be explored by examining the results of the following analyses: CFA, reliability, MTMM, and DTF.

The construct was evaluated by the CFA analysis and thus it becomes crucial to explore the ability of the results of the CFA to generalise about the construct being assessed. The factors identified in the CFA were primarily based on the EFA analysis previously conducted. This confirmed the existence of an inherent structure and underlying construct of the ECT. This was informed by the model fit statistics (chi-square, TLI, CFI, RMSEA, RMR, AIC, and CAIC), which all indicated that the hypothesised model was acceptable and fitted the model. Additionally, the variance explained by different factors as well as the relationships of the factors to the model substantiated this claim. This confirms that the underlying construct of verbal reasoning was present in both the test versions and was comprised of a few factors, as indicated in the CFA analysis. Based on these CFA results, the ability to generalise about verbal reasoning is not only tied to the test but rather suggests that the broader construct of verbal reasoning can be considered.

The reliability coefficients across test versions remained consistent even when improved by deleting items. Moreover, the improved reliability coefficients were acceptable for aptitude assessments, implying that the construct of the instrument can be relied on as inferences can be made. These inferences are tied to the construct being assessed as verbal reasoning and a good reliability value allows one some certainty in the consistency of the assessment of this construct. Based on this, the argument for generalising about the construct of verbal reasoning is influenced by the good reliability (improved reliability coefficient that is suitable for aptitude tests) and its internal consistency regarding the construct.

The results of the MTMM were of particular interest as the construct of verbal reasoning was assessed by exploring various correlations. The correlations observed between the ECT and the psychometric tests (Verbal Reasoning, Vocabulary, Reading Comprehension, and Long-Term Memory) that were theorised to be connected to the ECT were related and shared the strongest relationships. This was observed for both test versions, confirming that these relationships are consistent. These correlations, therefore, indicate that the construct being assessed (verbal reasoning), were present across test versions and consisted of different factors (which were consistently observed across the test versions). This relationship acknowledges the existence of the underling construct of verbal reasoning as well as the factors that were involved. The existence of the construct is important, especially when being assessed by external sources of information. This furthers the argument that verbal reasoning is being assessed by the ECT, but is not limited in terms of its interpretation and generalisability.

The gender DTF results indicated that the fit statistics for gender were similar for both test versions and proposed that the persons of the sample were problematic in terms of their limited variation of abilities across test versions. The items for the different genders were acceptable in terms of their variation of difficulties and improved across test versions. The gender fit statistics also revealed a high reliability and good average infit and outfit MNSQ values, which creates some certainty of the fact that the gender performance was not necessarily biased and they performed similarly on the items. It does, however, imply that the persons' limited ability levels meant they were not able to perform better, but the items were not the cause of any specific bias linked to ability. There were only a few items identified as problematic, but they were not the same across the test versions.

Based on this information, the DTF results for gender suggest that when removing the problematic items from the respective test versions, the remaining items do not function differently for the different genders. This implies that there is no invariance for the remaining items, as the individuals perform similarly. This allows one to generalise across gender groups with regards to their verbal reasoning ability.

The DTF results for the different racial groups (African, White, and Coloured) indicated a similarity across the three racial groups for the two test versions, as similar issues were observed. The persons of the sample for the three racial groups were problematic in terms of their variation of abilities and their reliability value could be higher. The items improved in terms of their variation of difficulties across the three racial groups and a high reliability and good average infit and outfit MNSQ values were observed. There were quite a number of problematic items identified for the combinations of racial groups (African and White; African and Coloured).

Based on these items, it becomes clear that when considering the performance of difference racial groups on the items, some racial groups seem to be favoured. This suggests that there were persons who were invariant for particular items of the ECT. Since there are quite a number of items for both test versions on which persons were invariant, the ability to generalise across racial groups becomes challenging. The majority of the items, however, allow for persons to perform similarly regardless of race. Some items have nonetheless been identified as problematic. This consequently limits the generalisability of the ECT across racial groups.

The argument regarding the generalisability facet of construct validity was assessed by evaluating the results of the CFA, reliability, MTMM, and DTF. These results provide a complete picture of the verbal reasoning construct and its ability to be generalised beyond the confines of the test. It is therefore evident that the CFA, reliability, and MTMM results support the generalising of the construct of verbal reasoning. The DTF results on the other hand, highlighted problematic items for gender and particularly for the different race groups. Thus, the generalisability of verbal reasoning as a construct is limited to the removal or exclusion of these problematic items, as they will be biased and provide a distorted view of the construct.

8.7.5 External Facet of Construct Validity

The external facet of construct validity concerns the construct of the test and requires one to provide external supporting evidence for the existence of the construct. These measures must, however, be known to measure the hypothesised construct of the test. The external measures therefore confirm the meaning of the construct across the test and other tests. The external facet of construct validity can be established by conducting MTMM analyses. This allows one to assess whether the same construct is present across similar tests, as they both would claim to measure it. This would assist in establishing convergent and discriminant validity, thereby confirming that the specified construct is indeed being measured (Baghaei & Amrahi, 2011; Messick, 1995, 1996; Ravand & Firoozi, 2016; Smith, 2001). The external facet of construct validity can therefore only be established by exploring the results of the MTMM conducted on the ECT for both test versions (Ravand & Firoozi, 2016).

When examining the various correlations observed for both test versions, particular relationships were observed for the two test versions that need to be considered to establish if the external facet of construct validity was observed. It should be noted that the ECT is theorised to be measuring verbal reasoning and for this reason, the correlations that one expects to be connected to the ECT are also related to the construct of verbal reasoning. The psychometric tests that were theorised to be more closely related to the ECT were related. These psychometric tests were Verbal Reasoning, Vocabulary, Reading Comprehension, and Long-Term Memory. Within the literature section (section 4.2, 4.3 and 4.4), it was noted that these constructs (the psychometric tests identified) are closely related due to the shared reasoning element. Additionally, the ECT and these psychometric tests (those theoretically related to the ECT) shared the strongest relationships when compared to the correlations between the psychometric tests that were not theorised to be related.

The consideration of the relationship observed between the ECT and these constructs as well as the relationship observed between these constructs among themselves are important. This allows one to further the argument of construct validity as the construct is observed in multiple correlations. Since these correlations also confirmed the existence of the theorised relationship, one can confirm the theorised existence of verbal reasoning as the core construct of the ECT. Furthermore, within this construct, several other related constructs such as vocabulary are present. Since these relationships were observed across test versions, it would suggest that an inherent construct is present in the ECT.

The relationship observed between the ECT and the psychometric tests that were hypothesised to be less or not related to the ECT were: Non-Verbal Reasoning, Numeric Comprehension, Mechanical Insight, Calculations, Comparison, Pattern Completion, Figure Series, Spatial 2D, and Spatial 3D. There was a small relationship observed between the ECT and the constructs mentioned. Since these correlations were relatively small, the rationale behind this small relationship is possibly due to the reasoning component that is common to all these constructs. This allows one to consider the fact that the various forms of reasoning are interrelated. This was a common theme for both test versions.

Based on the relationships observed in the MTMM analyses for both test versions, the correlations suggest that the ECT is definitely tapping into a reasoning construct. The relationship with the theorised psychometric tests further supports the claim that this construct is possibly verbal reasoning, as the relationships are indicative of reasoning which is largely represented in a verbal way. This, therefore, allows one to argue that the external facet of construct validity has been met as the strongest relationships observed are indicative of the hypothesised construct.

8.7.6 Consequential Facet of Construct Validity

The consequential facet of construct validity requires one to explore intended and unintended results that can occur from the use of the test and testing individuals. This includes factors that affect how individuals perform on the test as well as adverse effects that arise from the testing. The consequential facet of construct validity can be established by exploring DIF and the item-person map results and interpretations. This allows one to investigate whether there are issues that may influence persons based on the test items. The decisions made about problematic items are also important as they may have consequences for testing and the population at large (Baghaei & Amrahi, 2011; Messick, 1995, 1996; Ravand & Firoozi, 2016; Smith, 2001). For this reason, the consequential facet of construct validity was assessed by exploring the results of the Rasch analysis and DTF, as these analyses were most suited to investigate the concerns associated with testing (Baghaei & Amrahi, 2011; Ravand & Firoozi, 2016; Smith, 2001).

In the Rasch analysis, the item-person map is of specific focus as it compares the persons to the items on a continuum. This continuum provides a visual representation of the relationship between the persons and items in terms of ability (Ravand & Firoozi, 2016). The results for both test versions were similar as they both indicated that the test was generally well targeted for the sample of individuals. This deduction is made from the evidence that most items catered to the ability levels of the individuals. It was, however, evident that there were items that were higher than the individual's ability (observed in the ECT version 1.3) as well as items that were lower than the individual's lowest ability (observed in the ECT version 1.2 and 1.3). This means that there are items that are too difficult and too easy for the sample used.

There were a few gaps in terms of the items and the persons, as some items had no corresponding person ability and some persons had no corresponding item ability. This was observed for both test versions. The item spread was, however, much better than the persons spread of ability as there are less ability levels covered by the persons compared to the items. There were also items that appeared redundant, as they were assessing the same difficulty level. These issues were present in both test versions. These results therefore suggest that most individuals would be able to answer the items of the ECT, except for the items that are of a higher difficulty level. Since the items appear to be catering to a variety of difficulty levels, some assurance is provided that low ability individuals will not be negatively affected, as the easiest item of the test was below the lowest ability individual. The spread of items seems to indicate that there are no particular biases identified by the items towards the persons, as the item-person map is relatively well spread. The consequential facet of construct validity has therefore been met when assessing the item-person map of the Rasch analysis.

The DTF results are essential when assessing for adverse effects resulting from testing. The DTF results considered both the fit statistics and the scatterplot results to ascertain if there were problematic items observed. The fit indices for gender (male and female) were similar for both test versions and suggested that the samples were problematic in terms of their limited variation of abilities across test versions. The items for the different genders were acceptable in terms of their variation of difficulties and improved across test versions. The gender fit statistics also revealed a high reliability and good average infit and outfit MNSQ values, which creates certainty of the fact that the gender performance was not necessarily biased and they performed similarly on the items. It does, however, propose that the person's ability levels limited their performance, but the items were not the cause of bias linked to ability. It should also be noted that according to the scatterplot and *t* statistics, only a few items were identified as problematic. These items were not all the same across the test versions.

Based on this information, the gender DTF results suggest that minimal adverse results would occur for different genders completing the test. Although a few items were problematic across the different genders for the test versions, the majority of the items were completed by both genders and did not indicate issues when observed with the scatterplot and *t* statistic. This therefore serves to argue that the consequential facet of construct validity has been met based on the DTF results for both genders as well as the Rasch analysis results obtained for both test versions.

The DTF results for the different racial groups (African, White, and Coloured) involved the evaluation of the fit statistics as well as the scatterplot and *t* statistic. These results are especially important when considering the political arena in which South Africa is located. The fit statistics of the Rasch analysis indicated that there was a similarity observed across the three racial groups for the two test versions, as similar issues were identified. The persons of the sample for the three racial groups were problematic in terms of their variation of abilities and their reliability value could be higher. The items improved in terms of their variation of difficulties across the three racial groups. In addition to this, a high reliability and good average infit and outfit MNSQ values were observed for the three race groups. There were quite a number of problematic items identified for the combination of race groups (African and White; African and Coloured). There was also nine common items identified among these race groups that were problematic.

It becomes clear that when considering the performance of difference race groups on the items, some races seem to be favoured based on the items. It should, however, also be noted that although there were quite a number of problematic items identified among the racial combinations, the majority of the items were not problematic across racial groups. When assessing this and considering the political nature of South Africa, these racial discrepancies become contentious. Bearing this in mind, the results of the DTF for the three racial groups suggest that the consequential facet of construct validity has partially been met.

In the discussion of the Rasch and DTF results regarding gender and the different racial groups, it becomes apparent that there is evidence of most items adhering to acceptable standards. The acknowledgement of problematic items, however, suggests that there are items that need serious consideration when considering the sample involved, particularly the racial groups. Based on these results, it is fair to state that the majority of the items across test versions can be used without causing adverse effects. The test as a whole, however, needs to acknowledge the possibility of biased items and may need revision to avoid unfair discrimination and favouring particular race groups.

The results of the five objectives were interpreted using Messick's unified theory of construct validity to assess whether construct validity was achieved. Within Messick's theory, the six facets of construct validity were examined and argued according to the statistical analyses conducted for the five objectives. The findings of the various analyses suggested that four of the six facets (content, structural, external, and generalisability) of construct validity were met. The substantive and consequential facets of construct validity could be only partially met when considering all the implications of the results. Construct validity was however achieved as most of these aspects were met and the partially met facets would be cautioned against.

8.8 The Construct of Verbal Reasoning

This paper was focused on establishing the construct validity of the ECT and as a result, the construct of verbal reasoning was one of the aspects that required inspection. When observing the literature regarding verbal reasoning (Chapter 4), there is limited information (Holyoak, 2012; Roomaney & Koch, 2013; Strand, 2004) that explicitly indicates what exactly verbal reasoning is and how precisely one should measure it. In the literature, the broad term of reasoning is defined in terms of inferences and logical thinking, which is expressed as either deductive or inductive reasoning. There is also mention of a strong relationship between these two methods of reasoning (Lohman & Lakin, 2009). When considering the nature of the ECT and the factors that emerged, there is clear evidence that both deductive and inductive reasoning was required of individuals completing the ECT. Deductive reasoning was identified in the items that required individuals to infer from the comprehension they had read. Inductive reasoning was also observed in the items and was based on the items in the language section of the ECT. Both these types of items formed part of the reasoning and deduction factors identified in the ECT.

The information obtained from the CFA conducted on the ECT version 1.3 yielded interesting results and indicates to some extent that there is a relationship between deductive and inductive forms of reasoning. In terms of the CFA, the following was worth noting. Firstly, the reasoning factor was positively skewed, while the deduction factor was negatively skewed. This implies that the reasoning factor was experienced as very difficult by individuals while the deduction factor was experienced as very easy. Secondly, the reasoning factor (0.409) had the second highest correlation estimate and the deduction factor (0.289) factor had the third highest correlation estimates in accounting for the variance in the model. Thirdly, the reasoning factor (0.123) had the second highest loading on verbal reasoning (underlying factor of the ECT), while the deduction factor (0.447) had the smallest loading on verbal reasoning. Fourthly, the reasoning factor had the second highest effect on verbal reasoning and it explained 64% of the variance. The deduction factor had the third largest effect on verbal reasoning and it explained 54% of the variance. These findings therefore suggest that these factors of verbal reasoning are important in defining and constructing verbal reasoning.

Linking to these findings, the mental rules theory (Lohman & Lakin, 2009; Manktelow & Chung, 2004) is also applicable. This theory highlighted the various aspects of reasoning identified in the compilation of the factors of the ECT as well as the correlations observed with other psychometric assessment constructs. More so, this theory identified strongly with deductive reasoning and recognised errors in reasoning to be linked to working memory ability (Lohman & Lakin, 2009). It should be noted that there were small correlations observed for the ECT and long and short-term memory. This means that this theory is still relevant and applicable with regards to defining reasoning. This also links to Jensen’s (1974) claim that intelligence tests include memory. All reasoning constructs, such as verbal reasoning and non-verbal reasoning, were observed through the relationship between the ECT and these constructs (MTMM results).

Although these findings are highly valuable in terms of confirming the construct of reasoning and specifically verbal reasoning in the ECT, they do, however, leave room for a particular concern. This unease can be expressed as the issues related to the testing methodology used. The phrasing and the type of information elicited from the verbal reasoning tests (DAT 2 and AAT 2) highlighted this concern. The similarity across these two tests, which were adapted for South Africa and based on Western theory, included answering patterns with the phrases “is to” and “as”. For example: “Chick is to baby as hen is to mother”. This is an analogy type of phrasing where the individual answering the question must make the link between the two objects proposed. This is generally not problematic, but this is not a commonly used phrase in South Africa’s education system (Koch, 2015; Roomaney & Koch, 2013).

This answering pattern concern was observed by the researcher when assisting in selections that contained cognitive assessments such as the DAT 2 and AAT 2. Connecting with this is the notion that language is the method through which individuals make sense of things. It influences how one thinks as well as impacts how one understand things, thereby allowing individuals to differ in terms of cognition. An individual’s unfamiliarity with test content may also influence their performance on cognitive assessment (Boroditsky, 2011; Malda et al., 2010). This legitimises the concern that analogies may be hindering cognitive performance of South Africans, especially African individuals, as they are multilingual and multicultural people (Koch, 2015; Roomaney & Koch, 2013).

In addition to this, the study on the verbal analogies subscale of the Woodcock Munoz Language Survey, which was adapted for use in South Africa and needed to be adapted for the Xhosa population, provided valuable insight on the understanding of verbal analogies in South Africa. The research on this subscale indicated that it was particularly problematic to adapt and a new subscale that would be more culturally appropriate had to be developed (Koch, 2009, 2015). This verbal analogies scale was, however, problematic as it contained biased items and equivalence was not established (Koch, 2015; Roomaney & Koch, 2013).

Thus when examining the ECT, the researcher initially expected it to be linked to constructs such as reading comprehension and vocabulary, but based on the readings (Chapter 3 and 4), the researcher began to theorise that it could possibly be measuring verbal reasoning. It is natural to assume that when using language that it relates to only the theory of language but this is flawed. The mind uses language as a vehicle and it requires reasoning and perception to do so. Thus, language is a combination of language-related structures and cognitive processes. These languages structures require cognitive thought (Radden, 2008). Moreover, reading texts requires the use of memory, which assists in generating meaning and allowing inferences to be made. This is also essentially how comprehension operates and language comprehension is therefore linked to cognition because of the cognitive tasks required to create understanding and coherence (Gernsbacher, 1990; Pretorius, 2002; Van den Broek & Gustafson, 1999). Thus, the link between language and cognition is inevitable and ever present.

The overarching notion that the paper intended to explore was that the ECT measures verbal reasoning (mentioned in Chapter 1). This could, however, not merely be deduced by literature, but rather by a combination of literature and statistical analyses. This would make it not only theoretical but also practically plausible. For this reason, the construct validity and reliability was explored to ascertain that the construct of the ECT was, in fact, valid. The discussion of the two test versions highlighted that both tests produced fairly similar results in terms of the items, construct, and reliability. Additionally, the theoretical framework used, Messick’s unified theory of construct validity, allowed the various analyses to be interpreted and construct validity was preliminarily established.

The results of the MTMM analyses were of particular interest as they provided very insightful information about the construct. Based on these results, the construct of the ECT was linked to the following constructs: reading comprehension, vocabulary, verbal reasoning, long-term memory, short-term memory, non-verbal reasoning, mechanical insight, numeric comprehension, calculations, comparisons, pattern completion, figure series, Spatial 2D, and Spatial 3D. The range of its relationship with each of these constructs is, however, more important and is what separates the construct of the ECT from simply being considered a general reasoning test. The rationale for the ECT being related to these very diverse constructs lies in the fact that the reasoning component of the ECT is able to tap into these various components (Marshalek, 1981). Furthermore, the strongest relation for the ECT was with the following constructs: reading comprehension, vocabulary, verbal reasoning, long-term memory, and short-term memory (particularly the ECT 1.3). The commonality among these constructs is that they are verbally orientated. Moreover, the strongest relationship was between verbal reasoning and reading comprehension, which corresponded to literature (Lakin, 2012). The ECT displayed a similar bond with reading comprehension. The similarity between the verbal reasoning assessment and the ECT can therefore be confirmed.

The link between verbal ability tests and intelligence, specifically crystallized intelligence, was identified in literature when observing the relationship among intelligence tests. More so, these verbal tests were better at predicting general intelligence and showed a relationship to comprehension (Gignac, 2006; Horn & McArdle, 2007; Jensen, 1974; Kvist & Gustafsson, 2007; Marshalek, 1981). This was similarly observed in the relationships of the ECT across all the constructs of the DAT, SAT, and AAT tests. This confirms the finding that verbal ability tests generally relate to all forms of intelligence. This finding is important as it emphasises the wide spread of verbal ability. One can argue, to an extent, that verbal reasoning as a form of reasoning relates to most of these constructs.

The literature indicated that there is a relationship between vocabulary and reasoning, due to the required cognitive processes such as drawing inferences and comprehending (Lohman & Lakin, 2009; Marshalek, 1981). This relationship was observed in the CFA results, in that there was a relationship between the vocabulary factor and verbal reasoning. Within the CFA results, vocabulary was one of the factors that contributed to the construct of verbal reasoning and explained some of the variance in the model. Additionally, the vocabulary factor explained 48% of the variance in verbal reasoning. For the MTMM results, there were high correlations between Vocabulary and Verbal Reasoning (0.616 in ECT version 1.2 and 0.536 in ECT version 1.3) and between Vocabulary and the ECT (0.729 for ECT version 1.2 and 0.633 for ECT version 1.3). The implications of the relationship observed for the ECT, particularly the ECT 1.3, create an argument for verbal reasoning.

The psychological construct of verbal reasoning in the ECT is therefore complex and is not demonstrated as most verbal reasoning tests, but rather as a South African version of verbal reasoning. Moreover, one can view the traditional concept of verbal reasoning as a colonised concept and thus the construct of the ECT presents a decolonised concept that is more psychologically appropriate for the South African population. The psychological nature of this decolonised concept of verbal reasoning is similar to the traditional concept of verbal reasoning in that it relates to all aspects of reasoning in other constructs and identifies strongly with the verbal components of reasoning.

The decolonised version of verbal reasoning therefore signifies the importance of establishing tests within South Africa’s multicultural context. This version of verbal reasoning was partially observed in the results of the DTF, as the majority of the items were unbiased across racial and gender groups. These results provide evidence that the majority of the items are appropriate, although there are items that require attention to eliminate biases.

These findings also confirm the notion that culture has a great influence on the items used in the test as well as the evaluation of the test (Davies, 2003).

The awareness and investigation of cultural factors are vital when evaluating the construct validity of a test as well as guiding the development of a test to ensure that it is culturally appropriate for the population. By addressing these cultural issues, one reduces the effects of cross-cultural bias that confront all psychometric assessments. Since the ECT was evaluated using statistical analyses and the results were interpreted by an individual familiar with the culture, the effects of culture can be considered as appropriately dealt with (Mushquash & Bova, 2007; Tseng, 2001; Van de Vijver & Rothmann, 2004).

The identification of verbal reasoning in this paper can be regarded as a paradigm shift and the establishment of a new model. This shift in thinking is how psychological research on verbal reasoning and test development in psychology can be advanced. The conceding of a decolonised notion of verbal reasoning can be viewed as a form of social justice in that the inequality that often plagues cognitive assessment was responded to This psychological inquisition granted an opportunity to create a new definition of verbal reasoning, which is the composition of deductive and inductive reasoning skills to identify plausible responses to verbal stimuli. This definition that was proposed in Chapter 4 has been theoretically and practically verified, in terms of the literature explored and the relationships observed in the statistical analyses.

8.9 Conclusion

The comparison of the two test versions for the various statistical techniques was necessary as it highlighted the improvement across the test versions. It also identified the core elements of the two test versions and made some reference to literature. This allowed the results to appear not merely statistical, but also psychologically relatable.

The six facets of construct validity were thoroughly explored and argued according to the statistical analyses that contributed to the facet being investigated. The results became more relevant and had more impact when reviewed from the theoretical viewpoint of these six facets. The importance of sufficiently arguing for construct validity using these analyses allowed one to identify the significant implication of such findings. Moreover, the consistency across test versions served to promote the relevance of each statistical technique used. The findings, therefore, suggest that four of the six facets (content, structural, external, and generalizability) of construct validity were met according to the evidence provided by the various analyses. The substantive and consequential facets of construct validity could only partially be met when considering all the implications of the results. This, however, provides certainty regarding the construct of verbal reasoning and the overwhelming evidence demonstrates that the ECT is indeed measuring verbal reasoning.

One of the most important aspects of this study was the argument regarding verbal reasoning. It was only stated later in the paper, as the statistical aspects and theoretical ground needed to be laid for the argument to be substantial. The notion that the ECT is measuring verbal reasoning was argued not only on a statistical level but also on a theoretical level. The more profound revelation was, however, that the ECT is not measuring the conventional notion of verbal reasoning but rather a deconstructed and decolonised version of the concept. This version would be more appropriate for the South African context and would be less problematic when compared to the verbal analogies used in verbal reasoning assessments.

There was, therefore, sufficient evidence to suggest that verbal reasoning was observed across test versions. Additionally, the reliability of the ECT suggests that when a few problematic items are removed, the ECT is sufficiently reliable to measure and report on verbal reasoning as a construct. This addresses the two core aims of this study: The construct validity and reliability of the ECT was observed.

CHAPTER 9: RECOMMENDATIONS, LIMITATIONS, AND CONCLUSION

9.1 Introduction

The study employed several analyses such as the Rasch analyses, confirmatory factor analysis (CFA), multi-trait multi-method (MTMM), differential test functioning (DTF), and reliability to address the core aims of the study. These aims were to explore the construct validity and reliability of the ECT. The statistical analyses were instrumental in establishing whether the items and dimensions of the test exhibited the construct. After discussing the implications of the findings in the previous chapter, the need to assess the possible shortcomings associated with the study are essential as they limit the application of the findings. Moreover, the findings and discussion suggest that additional research needs to be conducted to further the understanding of verbal reasoning and provide more certainty concerning the validation of the ECT. This chapter will therefore explore the recommendations and limitations of this study as well as conclude the study.

9.2 Recommendations

When validating a test such as the ECT, several aspects need to be considered, as this test is in a process of development and requires refinement. This implies that careful consideration needs to be given to the ECT, as specific elements need to be addressed before one can use the test with certainty. These elements include the construct, the items, and the application of the test. These elements combined have various consequences and it is the developer’s responsibility to ensure that these consequences are minimal and result in a correct decision.

For this study, the two test versions were compared to see if there were improvements observed across these test versions. This comparison needed to be made for the various psychometric properties to be sufficiently explored. The comparison of the various statistical analyses performed across test versions evidenced that the ECT version 1.3 was indeed an improvement from the ECT version 1.2. This was confirmed by the item functioning and the observation of the construct. The MTMM, DTF, and reliability indicated similar results across the test versions and for this reason; one can argue that the consistency is also a progressive finding.

Bearing this in mind, it becomes fitting that any recommendations made will only be focused on the ECT version 1.3. The ECT version 1.2 will, therefore, not be used again for future research purposes as it has only served as a comparison to the ECT version 1.3. The recommendations will consequently be based on the various analyses conducted and will address the issues that arose for the ECT version 1.3 that need to be addressed by another study.

The Rasch analyses results for the ECT version 1.3 have several aspects that require further attention. Within the analyses, the very difficult items were identified as items 23, 25, 39, 40, and 42, while the easiest items were items 20, 22, 29, and 30. In addition to these items, it is clear that items 6, 9, 23, 30, and 40 were previously identified as outliers in the analyses conducted. The following items are therefore identified as problematic and need further investigation: items 6, 8, 9, 20, 23, 29, 30, 39, and 40. These items affected how individuals performed on the test and should be investigated further to assess for any bias. It would then be essential to perform differential item functioning (DIF) on each of these items, as they have appeared problematic and possibly need revision or to be removed. A DIF analyses will inform one of the possible reasons for these items performing in such a varied way.

Additionally, there was a serious issue identified within the analyses regarding the persons. The persons of the ECT had serious misfits. For the sake of comparison, these persons were included in the analyses for the ECT test versions 1.2 and 1.3. This is, however, an aspect worth exploring further, as the removal of misfitting persons (those observed in the response patterns) from the analysis and recalibration of the test may provide much better results for the ECT. This should, therefore, be explored in further research concerning the ECT.

The DTF results are of particular significance and necessitate future analyses in terms of the items identified for the different genders and race groups. In terms of the gender results, six problematic items were identified for the ECT version 1.3, namely items 4, 6, 26, 27, 32, and 36. The common problematic item identified for both the ECT version 1.2 and ECT version 1.3 was item 4 (true or false item). Thus, item 4 specifically needs to be explored, as it is a recurring item in terms of its problematic use across genders. Item 6 is a concerning item, as it was identified in both the Rasch analyses results and the DTF results as problematic. This would require investigation into why the item is problematic. The item would then need to be edited or removed from the test. Items 26 and 27 were identified as bottom items in the Rasch analyses results and are possibly too easy, but are not easy across genders. This would need to be investigated further. Items 32 and 36 did not appear in the Rasch analyses results, but would require investigation. All these items identified in the DTF analyses for gender would require a DIF analysis to be conducted. This will allow one to ascertain whether these items are, in fact, biased.

The DTF results identified 18 problematic items for the ECT version 1.3 between the African and White Group. These were items 2, 6, 7, 8, 9, 10, 11, 13, 18, 21, 24, 25, 26, 28, 32, 33, 39, and 40. There were nine common items identified for both the ECT version 1.2 and ECT version 1.3. These were items 2, 6, 9, 10, 11, 21, 28, 39, and 40. There needs to be an investigation into the content of these common items, as they seem to allow for differential performance based on the racial grouping.

The DTF results identified 14 problematic items for the ECT version 1.3 between the African and Coloured Group. These were items 3, 6, 7, 9, 10, 11, 18, 19, 21, 26, 28, 32, 33, and 42. There were seven common items identified for both the ECT version 1.2 and ECT version 1.3. These were items 6, 9, 10, 11, 18, 21, and 28. These mutual items require that their content be explored, as they seem to cause a discrepancy in the performance of the two racial groupings.

Items 6, 9, 10, 11, 21, and 28 were common across the racial groups as causing discrepancy in performance. Furthermore, items 6, 26 and 32 were identified in the DTF across gender and racial groups and indicate that there is a problematic pattern observed for cross-cultural testing. Based on the large number of items identified, a DIF analysis and content analysis should be conducted to understand the possible explanations for the differential performance across these two racial groups.

When comparing all the DTF results (for both gender and racial groups) in terms of the problematic items identified, items 6, 7, 9, 10, 11, 18, 21, 26, 28, 32, and 33 were common. These items are thus of utmost concern and would require investigation. The DIF results would inform whether they would need to be revised or removed.

The dimensionality information of the ECT, specifically the ECT version 1.3 from the Rasch analyses, it becomes apparent that there is possibly more information about the factor structure that can be attained. It would then be recommended that the following factor structures of the ECT version 1.3 be assessed. Firstly, an EFA should be run on all the problematic items identified in the Rasch analyses and be compared to the current factor structure. This will allow one to determine whether these items negatively impact the factor structure and variance explained by the model.

Secondly, the factor structures across gender and racial groups may provide very insightful information, particularly in terms of how the factors are assembled. This would require an EFA to be run. This may provide vital insights in terms of cross-cultural testing and research. There are more sets of analyses that can be run regarding the factor structure, but these recommendations given are most important and should be considered first in future studies on the ECT.

The reliability results were positive in that they demonstrated that the test was appropriate for research purposes and when items 6, 8, 9, 10, 11, 18, 23, and 25 were deleted, the *p* value improved. These items were also identified in other analyses (DTF and Rasch analyses) and emphasises the need for revision and an investigation into these items, as they not only influence how individuals perform but also lower the internal consistency of the test. Since they have been found across analyses, there is a serious need to explore these items.

In terms of the MTMM results, the correlations observed were important and because of the variety of psychometric tests used, the consideration of the constructs related to verbal reasoning became very pronounced. These results, however, do not serve to improve the development of the ECT. The relationship between the construct of verbal reasoning in the ECT and related and unrelated constructs of other psychometric assessments needs to be maintained with follow-up versions. The relationship between the verbal reasoning of the ECT needs to constantly be evaluated against other psychometric assessments as the development of the ECT proceeds. For this reason, the piloting of the ECT should, if possible, include the use of other psychometric tests by which the construct can be evaluated. This will ensure that all improvements also include improvements to the construct as a whole.

An additional aspect that requires attention and was only partially explored was the construct of verbal reasoning in a South African context. Further research that will provide more insight and increase local literature on verbal reasoning is required.

The test content and structure of the ECT requires some revising, which will improve the test. Additionally, the comprehension piece used in the test was used as an original piece and not edited, but this requires revision as it contains colloquial language which may be problematic for second language speakers. Based on the different aspects explored, the test will be improved and a more refined version will be created for further piloting to advance the test development and validation of the ECT.

9.3 Limitations

Several limitations are associated with this study and these limitations are grouped into technical and statistical issues. The technical concerns involve all the problems experienced by the researcher during the execution of the study. These relate to issues external to the study and some may have an influence on the results. The statistical issues identified are those concerns experienced when conducting the analyses. These concerns are internal to the study, as they had a direct impact on the results of the study.

Several technical issues were identified. Firstly, the ECT was empirically constructed and thus the constructs being measured are informed purely by statistical investigation and readings done by the researcher (developer). This could be a limitation as the focus is relatively broad, which can lead to multiple interpretations. Additionally, there was no expert guidance during the initial development or research conducted on the ECT version 1.2 and 1.3, which limits the information obtained. This has impacted on the quality of the test content which requires refinement. Secondly, the use of Excel sheets for data capturing as well as the use of “if” statements for answers may have led to errors. This occurs relatively easily, as the copying of formulas or even the copying of information from one Excel sheet to another can create discrepancies in the data. Although the data sheets have been checked several times, the reality of human error is possible.

Thirdly, the sample was conveniently selected and this leads to a restriction of range. These results can thus not be generalised and are specific to the population that was utilised. Fourthly, the item bank for the ECT is rather small, especially since it is still in development. This can be problematic when many items do not perform adequately in the test. Fifthly, the time limit imposed on the one version of the test could have impacted how the individuals performed. Their motivation, anxiety, and ability to complete a test within a specified time could have affected their performance on the test.

The statistical issues related to this study are based on the fact that the data limited the researcher's ability to suitably engage with the analyses. The use of secondary data have inherent limitations, as it does not allow one to conduct analyses easily as the data were already collected. The secondary data also restricts certain types of analyses to be completed. Most of the analyses were conducted with great ease, but the MTMM analyses were severely limited due to the lack of information available. The ability to conduct the traditional MTMM was not possible, and the researcher relied on correlations to assess the various constructs being measured. The analysis of the constructs did prove to provide useful information, but information that is more convincing may have been obtained if there were different sets of information available.

The exploration of verbal reasoning in the ECT provided a limited picture of this construct, as not all the aspects that comprise verbal reasoning could be measured by the ECT. This therefore restricts the construct of verbal reasoning to the aspects measured by the ECT. Thus, this construct of verbal reasoning is specific to the ECT and cannot be generalised.

9.4 Summary of Findings

This study embarked on evaluating the construct validity and reliability of the ECT. This required the use of various objectives to achieve these aims. These objectives are stated below and the summary of findings for each of these objectives will be mentioned. It should, however, be noted that upon the findings of these objectives, the construct of verbal reasoning, which was one of the aims of the study, was validated in the process of performing these objectives.

Objective 1: To statistically explore the unidimensionality of items using the Rasch model.

The findings of the Rasch analyses results were divided across the two test versions. These results yielded a wealth of information relating to the performance of both the persons and the items of the ECT. The results of the fit statistics revealed that the persons were consistently misfitting across the two test versions. They seem to be causing deviation in the analyses of the test. The item performance according to the fit statistics improved across the two test versions and only a few items deviated from the model. This was a very positive finding. Within the Rasch analyses results, the items and persons that deviated from the model were observed to assess where the deviation was occurring. This allowed the researcher to conclude that the persons were either careless, guessing, or not paying attention to the questions and thus irregular answering patterns occurred. The items causing differential performance were due to either difficulty or issues related to the item content. This, however, needs to be explored further. The dimensionality of the ECT was also explored through Rasch analyses and the results indicated that the test was multidimensional and there was some redundancy in the items and persons of the test, which limited the variance explained.

The variance explained by the model improved across the test versions. This is also a positive finding and suggests that the newer test version was an improvement from the earlier version.

Objective 2: To confirm the dimensionality of the ECT using CFA.

The dimensionality of the ECT was evaluated by CFA, which was done by conducting structural equation modelling (SEM) using SPSS AMOS. The model of the CFA for the ECT version 1.3 was mostly based on the results of the EFA conducted on the ECT version 1.2. This model was used to provide evidence in support of the construct validity of the ECT.

Objective 3: To support evidence of construct validity by conducting a MTMM analysis.

The MTMM analysis conducted did not conform to the traditional aspects that should be included due to the fact the required information was not available. Nonetheless, the core features of the traditional MTMM analyses were observed, as correlations with the various constructs were performed. These results were very interesting as they indicated that the strongest correlations with the ECT were observed among the following constructs: reading comprehension, vocabulary, verbal reasoning, and long-term memory. Moreover, the ECT was correlated with constructs that were not hypothesised to relate, such as calculations, mathematical comprehension, mechanical insight, spatial 2D, and spatial 3D. These correlations with constructs not hypothesized to relate emphasised the strongest component of the ECT, which is reasoning ability, and thus it would be expected that there be correlations across all constructs. These relationships were observed for both test versions and only differed in terms of the size of the relationship. This could, however, be attributed to the time limit imposed on the ECT version 1.2.

Objective 4: To explore the measurement invariance using DTF.

The measurement invariance was explored by conducting a DTF on both the test versions using Winsteps (Linacre, 2009). This created both a spread sheet and scatterplot of the different groups of data observed. For the DTF analyses, the gender and racial groups were compared. This was done for both test versions. Interestingly enough, the results were similar across the two test versions for both gender and racial groups, indicating a consistency across the test versions. The DTF results for the gender comparison revealed a few items that were possibly biased across the genders. The DTF results for the different racial groups (African and White; and African and Coloured) indicated more possibly biased items across the racial groups. These items that were identified are a cause of concern and require further investigation to establish what specifically they could be measuring. It is, however, worth noting that the majority of the items were considered appropriate for both genders and the different racial groups. This was observed across both test versions.

Objective 5: To evaluate the internal consistency of the ECT by conducting a reliability analysis.

The internal consistency was evaluated by the Kuder-Richardson Formula 20. The reliability coefficient indicated that both test versions were relatively reliable, as these test versions could be used for research purposes. When exploring the items in the item-total correlations, the items that were not adding value to the reliability coefficient were removed. This improved the internal consistency of both test versions and these test versions could then be used for assessing aptitude. This is important, as this test is intended for aptitude assessments and therefore shows great promise. It should, however, be noted that the items that posed a threat to the reliability of the test need to be revised as they are not adding value to the internal consistency of the ECT.

9.5 Conclusion

The consideration of the limitations of the study is important as it informs one of restrictions of the way in which the data were collected and as such, the way in which the information can be used. The limitations are vital when considering further studies, as they can either be avoided or handled better in future. They are, however instances that occur, either as part of the research process or as a hindrance to the information obtained.

The recommendations that were made are essential to the study, as the findings suggest that more research should be conducted. Due to the nature and objectives of this study, there are elements that still need to be explored and will assist in the further validation of the ECT. The recommendations suggested are, however, primarily based on the information obtained from the results. Thus, further studies will aid in producing a more psychometrically sound assessment of verbal reasoning.

When reviewing the summary of the findings, the statistical techniques used contributed to the development of the ECT, provided evidence of construct validity, and established the reliability of the ECT. The information provided by these techniques is vital to test development both internationally and in South Africa. The consideration of a multicultural and multilingual context makes developing a test not only a challenge but also a worthwhile venture. The philosophical discussion on verbal reasoning and the problematising of the traditional notion of verbal reasoning as a Euro-American was significant to this study. It allowed for a new discourse on the psychological construct of verbal reasoning, whereby a new system of thought was created. This new discourse framed the construct of verbal reasoning in the ECT as both deconstructed and decolonised. This multifaceted construct still taps into the same theoretical constructs as the traditional notion of verbal reasoning but avoids the use of analogies, which has been observed as problematic, especially in the South African population.

To conclude, this study was able to advance the research and exploration of the concept of verbal reasoning, specifically within a South African population. Furthermore, the argument on verbal reasoning will advance psychological research on verbal reasoning and test development in psychology, particularly with regards to the creation of new discourses in this field. This is, therefore, an indispensable stride towards validating the ECT as a measure of verbal reasoning.

REFERENCES

Abrahams, F., & Mauer, K. F. (1999). The comparability of the constructs of the Sixteen Personality Factor in the South African context. *Journal of Industrial Psychology, 25*, 53-59.

Almeida, L. S., Prieto, M. D., Ferreira, A., Ferrando, M., Ferrandiz, C., Bermejo, R., & Hernandez, D. (2011). Structural invariance of multiple intelligences based on the level of execution. *Psicothema, 23*(4), 832-838.

Arbuckle, J. L. (2010). *IBM SPSS AMOS 19 users guide.* Chicago: AMOS Development Corporation.

Acton, G. S. (2003). What is good about Rasch measurement? *Rasch Transactions, 16*, 902-903.

Anderson, M. (1992). *Intelligence and development: A cognitive theory*. Oxford, United Kingdom: Blackwell.

Arce, C., De Francisco, C., & Arce, I. (2010). Multidimensional scaling: Concept and applications. *Papeles del Psicologo, 31*(1), 46-56.

Ardila, A., Ostrosky-Solis, F., & Bemal, B. (2006). Cognitive testing toward the future: The example of semantic verbal fluency (ANIMALS). *International Journal of Psychology, 41*(5), 324-332.

Arendse, D. E. (2010). *Evaluating the structural equivalence of the English and isiXhosaversion of the Woodcock Munoz Language Survey* (Unpublished master's thesis). University of the Western Cape, South Africa.

Arendse, D. E., & Maree, D. (2017). *Exploring the dimensionality of the English Comprehension Test.* Unpublished article, University of Pretoria, Pretoria, South Africa.

Auer, P., & Wei, L. (2007). *Handbook of multilingualism and multilingual communication*. Europe: De Gruyter Mouton.

Baghaei, P. (2008). The Rasch Model as a construct validation tool. *Rasch Measurement Transactions, 22*(1), 1145-1146.

Baghaei, P., & Amrahi, N. (2011). Validation of a multiple choice English vocabulary test with the Rasch Model. *Journal of Language Teaching and Research, 2*(5), 1052-1060.

Barona, A., Reynolds, C. R., & Chastain, R. (1984). A demographically based index of premorbid intelligence for the WAIS-R. *Journal of Consulting and Clinical Psychology, 52*(5), 885-887.

Bekwa, N. N. (2016). *The development and evaluation of Africanised items for multicultural cognitive assessment* (Unpublished doctoral dissertation). University of South Africa, Pretoria, South Africa.

Belenky, M. F.; Clinchy, B. M.; Goldberger, N. R.; Nancy, R. & Tarule, J. M. (1986). *Women's Ways of Knowing: The Development of Self, Voice and Mind*. New York: Basic Books.

Berk, L. E. (2006). Language development. In L. E. Berk, (Ed.), *Child development* (8th ed., pp. 356-395). Boston: Allyn & Bacon.

Berk, S. (2004). Acquisition of verb agreement when first language exposure is delayed. In A. Brugos, L. Micciulla, & C. E. Smith (Eds.), *BUCLD 28: Proceedings of the 28th* *annual Boston University Conference on Language Development* (pp. 62-73). Somerville, MA: Cascadilla Press.

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive tests: A UK meta analysis. *Journal of Occupational and Organisational Psychology, 78,* 387-409.

Betacourt, T. S., Yang, F., Bolton, P., & Normand, S. (2014). Developing an African youth psychosocial assessment: An application of item response theory. *International Journal of Methods in Psychiatric Research, 23*(2), 142-160.

Binet, A. & Simon, T. (1908). The Development of Intelligence in the Child. *L'Annee Psychologique,* 14, 1-94.

Blake, B., & Pope, T. (2008). Developmental psychology: Incorporating Piaget's and Vygotsky's theories in the classroom. *Journal of Cross-disciplinary Perspectives in Education, 1*(1), 59-67.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model: fundamental measurement in the human sciences*. Hillsdale, New Jersey: Erlbaum Associates.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Boroditsky, L. (2011). How language shapes thought. *Scientific American.* Retrieved from www.ScientificAmerican.com.

Borsboom, D., & Mellenbergh, G. J. (2004). Why psychometrics is not pathological. *Theory and Psychology, 14*, 105-120.

Boslaugh, S. (2007). *Secondary data sources for public health: A practical guide.* New York: Cambridge University Press.

Brown, R. E. (2016). Hebb and Cattell: The genesis of the theory of fluid and crystallized intelligence. *Frontiers in Human Neuroscience, 10*(606), 1-11.

Budd, R. (1998). The Anglicization of American personality test: Rejoiners. *Selection and Development Review, 14*(2), 13-14.

Busing, F. M. T. A., Commandeur, J. J. F., & Heiser, W. J. (1997). Proxscal: A multidimensional scaling program for individual differences scaling with constraints. In W. Bandilla, & F. Faulbaum (Eds.), *Softstat 1997: Advances in statistical software* (pp. 67-74). Stuttgart: Lucius & Lucius.

Byrne, B. M. (2010). *Structural equation modelling with Amos: Basic concepts, applications and programming* (2nd ed.). New York: Taylor and Francis

Camarata, S., & Woodcock, R. (2006). Sex differences in processing speed: Developmental effects in males and females. *Intelligence, 34*, 231-252.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi trait-multi method matrix. *Psychological Bulletin, 56*, 81-105.

Cattell, R. B. (1963). Theory of Fluid and Crystallized Intelligence: A Critical Experiment. *Journal of Educational Psychology*, 54, 1-22.

Ceci, S. J. (1990). *On intelligence - More or Less. A Bio-Ecological Treatise on Intellectual Development*. Englewood Cliffs, New Jersey: Prentice Hall.

Chen, M. J., & Chen, H. (1988). Concepts of intelligence: A comparison of Chinese graduates from Chinese and English schools in Hong Kong. *International Journal of Psychology, 23*(1988), 471-487.

Clark, E., Garner, M. K., & Brown, G. (1992). Components of analogical reasoning in a mild head injured populations. *Current Psychology: Research and Reviews, 11*(1), 21-35.

Cockcroft, K., Bloch, L., & Moolla, A. (2016). Assessing verbal functioning in South African school beginners from diverse socio-economic backgrounds: A comparison between verbal working memory and vocabulary measures. *Education as Change, 20*(1), 199-215.

Cohen, J. (1988). *Statistical power analysis for behaviour sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Cooperman, E. W. (1980). Field differentiation and intelligence. *The Journal of Psychology, 105*, 29-35.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10*(7), 1-9.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological testing. *Psychology Bulletin, 52*(4), 281.

Csapo, B. (1997). The development of inductive reasoning: Cross-sectional assessment in an educational context. *International Journal of Behaviour Development, 20*(4), 609-626.

Davies, A. (2003). Three heresies of language testing research. *Language Testing, 20*(4), 355-368.

De Abreu, P. M., Baldassi, M., Puglisi, M. L., & Befi-Lopes, D. M. (2013). Cross-linguistic and cross-cultural effects on verbal working memory and vocabulary: Testing language-minority children with an immigrant background. *Journal of Speech, Language and Hearing Research, 56(2)*, 630-642.

De Beer, M. (2004). Use of differential item functioning (DIF) analysis for bias analysis in test construction. *South African Journal of Industrial Psychology, 30*(4), 52-58.

De Beer, M. (2011). The role of the learning potential computerized adaptive test (LPCAT) in the vocational guidance assessment of adolescents. *Educational and Child Psychology, 28*(2), 114-129.

Division of Statistics & Scientific Computation. (2012). *Structural Equation Modeling using AMOS.* Retrieved from https://stat.utexas.edu/images/SSC/Site/AMOS_Tutorial.pdf.

Dunne, T., Long, C., Craig, T., & Venter, E. (2012). Meeting the requirements of both classroom-based and systematic assessment of mathematics proficiency: the potential of Rasch measurement theory. *Pythagoras*, 33(3), 1-19.

Els, M., & Andries, C. (2011). *Measuring psychosocial development: A Rasch analysis of the sentence completion test for youth.* Paper presented at the European Conference on Development Psychology: 23-27 August 2011, Bergen, Norway.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Erguven, M. (2014). Two approaches to psychometric process: Classical test theory and item response theory. *Journal of Education, 2*(2), 23-30.

Eysenck, H. J. (1973). *The Measurement of Intelligence*: Lancaster: Medical and Technical Publishing Company.

Fagan, J. F. (2000). A theory of intelligence as processing: Implications for society. *Psychology, Public, Policy and Law, 6*(1), 168-179.

Field, A. P. (2005). *Discovering statistics using SPSS*. United Kingdom: Sage Publishers Limited.

Field, A. P. (2009). *Discovering statistics using SPSS.* United Kingdom: Sage Publishers Limited.

Fisher, W. P. (1992). The cash value of reliability. *Rasch Measurement Transactions, 22*(1), 1158-1161.

Flanagan, D. P., & McGrew, K. S. (1997). Across-battery approach to assessing and interpreting cognitive abilities: Narrowing the gap between practice and cognitive science. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories tests, and issues* (pp. 314-325). New York, NY: The Guilford Press.

Flanagan, D.P. & Ortiz, S.O. (2001). *Essentials of cross-battery assessment.*

New York, NY: John Wiley & Sons, Inc.

Fouche, F. A., & Verwey, F. A. (1982). *Manual for the Senior Aptitude Test: 1978 Edition.* Pretoria: Human Sciences Research Council.

Foxcroft, C. D. (2004). Planning a psychological test in the multicultural South African context. *SA Journal of Industrial Psychology, 30*(4), 8-15.

Foxcroft, C. D., & Roodt, G. (2009). *An introduction to psychological assessment in South African context.* South Africa: Oxford University Press.

Foxcroft, C. D., Roodt, G., & Abrahams, F. (2013). Psychological assessments. A brief retrospective overview. In C. Foxcroft & G. Roodt (Eds.), *Introduction of psychological assessments in the South Africa context* (4th ed., pp. 9-27), Cape Town, South Africa: Oxford University Press.

Fraley, R. C., Waller, N. G., & Brennan, K. A. (2000). An item response theory analysis of English language learners: A critical discussion of select state practices. *Bilingual Research Journal, 29*(1), 31-42.

Galton, F. (1883). *Inquiries into Human Faculty and its Development*. London: Macmillan.

Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligences*. New York: Basic Books.

Garrison, M. S. (2009). The cognitive development of collegiate students: A brief literature review. *The Campbellsville Review*, 87-100. Retrieved from http://www.campbellsville.edu/websites/cu/images/library/campbellsville_review/vol_4/the _cognitive_development_of_collegiate_students_Garrison.pdf.

Garson, G. D. (2015). *Structural equation modelling.* North Carolina State, USA: G. David Garson and Statistical Associates Publishing.

Gentner, D., & Goldin-Meadow, S. (2003). Whither Whorf. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind. Advances in the study of language and thought.* Cambridge, MA: MIT Press.

Gernsbacher, M. A. (1990). *Language comprehension as structure building.* Hillsdale, NJ: Erlbaum.

Gignac, G. (2006). Evaluating subtest, ‘g’, saturation levels via the single trait correlated uniqueness (STCU) SEM approach: Evidence in favour of crystallized subtests as the test indicators of ‘g’. *Intelligence, 34,* 29-46.

Giri, R. A. (2003). Language Testing: Then and Now. *Journal of NELTA*, 8(1-2), 49-67.

Glass, G. (1976). Primary, secondary and meta-analysis of research. *Educational Research, 5*, 3-8.

Goldin-Meadow, S. (2007). Pointing sets the stage for learning language and creating language. *Child Development, 78*(3), 741-745.

Goodman, D. (2004). *A Multitrait-multimethod validity investigation of the 2002 Massachusetts Comprehensive Assessment System tests* (MCAS Validity Report No. 5). University of Massachusetts, Centre for Educational Assessment.

Green, K. E., & Frantom, C. G. (2002). *Survey development and validation with the Rasch Model.* Paper presented at the International Conference on Questionnaire Development, Evaluation & Testing, Charleston, SC. Retrieved from http://www.jpsm.undedu/qdet/final_pdf_papers/green.pdf.

Griffin, B., Carless, S., & Wilson, I. (2013). The undergraduate medical and health sciences admissions test: What is it measuring. *Medical Teacher, 35*, 724-730.

Griskevica, I., & Rascevska, M. (2009). The relationship among cognitive abilities and demographics factors in Latvia. *Baltic Journal of Psychology,* 55-72.

Guilford, J. P. (1956). The Structure of Intellect. *Psychological Bulletin*, 53, 267-293.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate data analysis.* United States of America: Prentice Hall.

Hamavandy, M., & Kiany, G. R. (2014). A historical overview on the concept of validity in language testing. *Advances in Language and Literacy Studies, 5*(4), 87-90.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational Measurement: Third Edition* (pp. 147-200). New York: Macmillan Publishing Company.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests. *Bulletin of the International Test Commission, 10,* 229-244.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practise, 12*(3), 38-47.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment.* United States of America: Lawrence Erlbaum Associates Limited.

Harris, C. L. (2006). *Language and cognition*. Massachusetts, USA: Boston University:

Harvey, R., & Hammer, A. (1999). Item Response Theory. *The Counseling Psychologist, 27*(3), 353-383.

He, J., & Van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture, 2*(2), 1-19.

Health Professions Council of South Africa (HPCSA). (1974). *Health Professions Act, Act 56 of 1974.* Pretoria, South Africa.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234-259). New York: Oxford University Press.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57*, 253-270.

Horn, J. L., & McArdle, J. J. (2007). Understanding human intelligence since Spearman. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical development and future directions* (pp. 205-247). Mahwah, NJ: Erlbaum.

Hoyle, R. (1995). *Structural equation modelling: Concepts, issues and applications.* Thousand Oaks, CA: Sage Publications.

Hu, L., & Bentler, P. M. (1999). Cut off criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modelling, 6*(11), 1-55.

Hula, W. D., Austermann-Hula, S. N., & Doyle, P. J. (2009). A preliminary evaluation of the reliability and validity of a self-reported communicative functioning item pool. *Aphasiology, 23*(7-8), 783-796.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge performance. *Journal of Vocational Behaviour*, 29, 340-362.

International Test Commission (ITC). (2000). *International guidelines for test use.* Retrieved from www.intestcom.org/itc projects.htm

Jaworska, N., & Chupetlovska-Anastasova, A. (2009). A review of Multidimensional Scaling (MDS) and its utility in various psychological domains. *Tutorials in Quantitative Methods for Psychology, 5*(1), 1-10.

Jensen, A. R. (1974). Interaction of Level I and Level II Abilities with Race and Socioeconomic Status. *Journal of Educational Psychology*, 66(1), 99-111.

Johnson-Laird, P. N. (2004). Mental models and reasoning. In J. P. Leighton & R. J. Sternberg (Eds.), *The nature of reasoning* (pp. 169-204). New York, NY: Cambridge University Press.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527-535.

Kane, M. T. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131-153). New York: Routledge.

Keeves, J. P. (1992). *The IEA study of science III: Changes in science education and achievement.* Oxford: Pergamon Press.

Keith, T. Z., & Reynolds, M. R. (2010). Cattell-Horn-Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools, 47*(7), 635-650.

Kendeou, P.; Van den Broek, P.; Helder, A. & Karlsson, J. (2014). A Cognitive View of Reading Comprehension: Implications for Reading Difficulties. *Learning Disabilities Research and Practise,* 29(1), 10-16.

Kennedy, S. W., Allaire, J. C., Gamaldo, A. A., & Whitfield, K. E. (2012). Race differences in intellectual control beliefs and cognitive functioning. *Experiential Aging Research, 38*, 247-264.

Kilfoil, W. R. (1999). The linguistic competence of science students. *South African Journal of Higher Education, 13*(1), 46-58.

Kline, R. B. (2011). *Principles and practise of structural equation modelling* (3rd ed.). New York, USA: The Guilford Press.

Koch, E. (2009). The case for bilingual language tests: A study of test adaptation and analysis. *Southern African Linguistics and Applied Language Studies (SALALS), 27*(3), 301-31.

Koch, E. (2015). Testing in bilingual education projects: Lessons learnt from the additive bilingual education project. *Per Linguam (A Journal for Language Learning), 31*(2). 79-93.

Koch, T. (2013). *Multilevel structural equation modeling multitrait-multimethod multi occasion data* (Doctoral dissertation). Retrieved from http//www.ewi.psy.faberlin.de.tkoch/…./index.html.

Koczwara, A., Patterson, F., Zibarrus, L., Kerrin, M., Irish, B., & Wilkinson, M. (2012). Evaluating cognitive ability, knowledge tests and situational judgement tests for postgraduate selection. *Medical Education, 46,* 399-408.

Kohut, M., Halama, P., Dockal, V., & Zitny, P. (2016). Gender differential item functioning in Slovak version of intelligence structure test 2000 revised. *Studia Psychologica, 58*(3), 238-250.

Kvist, A. V., & Gustafsson, J. (2007). The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell's Investment Theory. *Intelligence, 36*, 422-436.

Laher, S., & Cockcroft, K. (2013). *Psychological assessment in South Africa: Research applications.* Johannesburg: Wits University Press.

Lakin, J. M. (2012). Assessing the cognitive abilities of culturally and linguistically diverse students: Predictive validity of verbal, quantitative and non-verbal tests. *Psychology in the Schools, 49*(8), 756-768.

Lance, C. E., & Fan, Y. (2016). Convergence, admissibility and fit of alternative CFA models for MTMM data. *Educational and Psychological Measurement, 76*(3), 487-507.

Langdon, D. W., Rosenblatt, N., & Mellanby, J. H. (1998). Discrepancy poor verbal skills in poor readers: A failure of learning or ability. *British Journal of Psychology, 89*, 177-190.

Langer, D. A., Wood, J. J., Bergman, R. L., & Piacentini, J. C. (2010). A multitrait-multimethod analysis of the construct validity of child anxiety disorders in a clinical sample. *Child Psychiatry and Human Development, 41*(5), 549-561.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2009). *Winsteps version 3.68 computer software.* Retrieved from www.winsteps.com.

Linacre, J. M. (2011). *Winsteps tutorials 1: Analysis of dichotomous data.* Retrieved from www.winsteps.com/forum.

Linacre, J. M. (2012a). *Winsteps tutorials 1: Software operation and basic concepts.* Retrieved from www.winsteps.com/forum

Linacre, J. M. (2012b). *Winsteps tutorials 2: Fit analysis and measurement models.* Retrieved from www. winsteps.com/forum.

Linacre, J. M. (2012c). *Winsteps tutorials 3.* Retrieved from www.winsteps.com/forum.

Linacre, J. M. (2012d). *Winsteps tutorials 4.* Retrieved from www.winsteps.com/forum.

Linacre, J. M. (2015). *A user's guide to Winsteps Ministeps: Rasch model computer programs. Program manual 3.91.0.* Retrieved from www.winsteps.com.

Linacre, J. M. (2016). *Winsteps Rasch measurement computer program.* Retrieved from www.winsteps.com.

Linacre, J. M., & Wright, B. D. (2003). *A user's guide to BIGSTEPS Rasch model computer programs.* Retrieved from www.winsteps.com

Lindquist, K. A., & Gendron, M. (2013). What's in a word? Language constructs emotion perception. *Emotion Review, 5*, 66-71.

Little, R. J. A., & Rubin, D. A. (1987). *Statistical analysis with missing data.* New York: John Wiley & Sons.

Loehlin, J. C. (1997). *Latent variables models.* Mahwah, NJ: Lawrence Erlbaum Associates.

Lohman, D. F., & Lakin, J. M. (2009). Reasoning and intelligence. In R. J. Sternberg & S. B. Kaufman (Eds.), *Handbook of intelligence: Second edition* (pp. 1-47). New York: Cambridge University Press.

Long, C. (2011). *Mathematical, cognitive and didactic elements of the multiplicative conceptual field investigated within a Rasch assessment and measurement framework.* (unpublished doctoral dissertation). University of Cape Town, Cape Town, South Africa.

Long, C., Bansilal, S., & Debba, R. (2014). An investigation of mathematical literacy assessment supported by an application of Rasch measurement. *Pythagoras*, 35(1), 1-.16.

Maas, C. J. M., Lensvelt-Mulders, G. J. L. M., & Hox, J. J. (2009). A multilevel multitrait-multimethod analysis. *Methodology*, *5*(3), 72-77.

Mahoney, K. S., & MacSwan, J. (2005). Re-examining identification and reclassification of English language learners: A critical discussion of select state practices. *Bilingual Research Journal*, *29*(1), 31-42.

Malda, M.; Van de Vijver, F. J. R. & Temane, Q. M. (2010). Rugby versus Soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence*, 38(2010), 582-595.

Manktelow, K., & Chung, M. C. (2004). *Psychology of reasoning: Theoretical and historical perspectives.* New York: Psychology Press (Taylor & Francis Group).

Marais, A. C. (2007). *Using the Differential Aptitude Test to estimate intelligence and scholastic achievement of grade nine level* (Unpublished master's thesis). UNISA, South Africa.

Maree, D. J. F. (2004a). *Basic IRT tutorial for analyzing data. Tutorial 1 – The one parameter model or Rasch model-process. Starting with software.* Personal Collection of D. J. F. Maree, University of Pretoria, Pretoria.

Maree, D. J. F. (2004b). *Basic IRT tutorial for analyzing data. Tutorial 2 - Interpreting output for dichotomous data.* Personal Collection of D. J. F. Maree, University of Pretoria, Pretoria.

Maree, D. J. F. (2004c). *Basic IRT tutorial for analyzing data. Tutorial 3 - Interpreting output for dichotomous data.* Personal Collection of D. J. F. Maree, University of Pretoria, Pretoria.

Martin, C. R., & Savage-McGlynn, E. (2013). A 'good practice' guide for the reporting of design and analysis for psychometric evaluation. *Journal of Reproductive and Infant Psychology, 31*(5), 449-455.

Marshalek, B. (1981). *Aptitude research report: Trait and process aspects of vocabulary knowledge and verbal ability* (Report No. 15). Stanford, CA: Stanford University, School of Education.

Marshalek, B., Lohman, D. F., & Snow, R. E. (1983). The complexity continuum in the Radex and hierarchical models of intelligence. *Intelligence*, *7*, 107-127.

McBride, N. L. (2001). *An item response theory analysis of the scales from the international personality item pool and the neo personality inventory-revised* (Unpublished master's dissertation). Virginia Technology University, Blacksburg, VA.

McDonald, E., & Van Eeden, R. (2014). The impact of home language on the understanding of the vocabulary used in the South African version of the Sixteen Personality Factor questionnaire fifth edition. *South African Journal of Psychology, 44*(2), 228-242.

Meiring, D., Van de Vijver, F., & Rothmann, S. (2006). Bias in the adapted version of the 15 FQ questionnaire in South Africa. *South African Journal for Psychology, 36*, 340-356.

Meiring, D., Van de Vijver, A. J. R., Rothmann, S., & Barrick, M. R. (2005). Construct, item and method bias of cognitive and personality tests in South Africa. *South African Journal of Industrial Psychology, 31*(1), 1 -8.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741- 749

Messick, S. (1996). Validity and Washback in language testing. *Language Testing 13*(3), 241-256.

Milani, T. M. (2007). Language testing and citizenship: A language ideological debate in Sweden. *Language in Society, 37*, 27-59.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 111-121.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory and Psychology, 10*, 639-667.

Michell, J. (2008). Conjoint Measurement and the Rasch Paradox: A Response to Kyngdon. *Theory and Psychology, 18(1)*, 119-124.

Morgan, B. (2015). *Introduction to confirmatory factor analysis and structural equation modelling using AMOS.* Personal Collection of B. Morgan, University of the Western Cape, Cape Town.

Moroke, N. D. (2014). Profiling some of the dire household debt determinants: A metric multidimensional scaling approach. *Journal of Economics & Behavioural Studies, 6*(11), 858-867.

Morrow, K. (1981). Basic concerns in test validations. In J. C. Alderson & A. Hughes (Eds.), *ELT documents III- Issues in language* (pp 9 – 25). London: The British Council.

Mupawase, A., & Broom, Y. (2010). Assessing cognitive linguistic abilities in South African adults living with HIV: the cognitive linguistic quick test. *African Journal of AIDS Research, 9*(2), 147-162.

Mushquash, C. J., & Bova, D. L. (2007). Cross-cultural assessment and measurement issues. *Journal of Development Disabilities, 13*(1), 55- 66.

Nell, V. (2000). *Cross-cultural neuropsychological assessment: Theory and practise.* New Jersey: Lawrence Erlbaum Associates.

Nunnaly, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw Hill.

Oberauer, K., & Lewandowsky, S. (2013). Evidence against decay in verbal working memory. *Journal of Experimental Psychology*, 142(2), 380- 411.

Oberholzer, B. (2005). The relationship between reading difficulties and academic personality factor in the South African context. *Journal of Industrial Psychology, 25*, 53-59.

Ormrod, J. E. (2008). *Cognitive development. Educational psychology, developing learners* (6th ed.). New Jersey: Pearson Education Inc.

Owen, K. (2000). *Manual for the Differential Aptitude Test: Form L*. Pretoria: Human Sciences Research Council.

Owen, R., & De Beer, J. F. (1997). *Manual for the Academic Aptitude Test (University).* Pretoria: Human Sciences Research Council.

Pae, H. K. (2011). Differential item functioning and unidimensionality in the Pearson Test of English Academic. *Pearson Education Ltd*, 1-6.

Pae, H. K. (2012). Construct validity of the Pearson Test of English Academic: A multitrait-multimethod approach. *Pearson*, 1-8.

Pae, H., Greenberg, D., & Morris, R. D. (2012). Construct validity and measurement invariance of the Peabody Picture Vocabulary Test III Form A in the performance of struggling adult readers: Rasch Modelling. *Lang Assess Q*, *9*(2), 152 -171.

Pae, T. I., & Park, G. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing*, 23(4), 475-496.

Pal, H. R., Pal, A., & Tourani, P. (2004). Theories of intelligence. *Everymans' Science*, *3*, 181-185.

Palmer, A. S. & Bachman, L. F. (1981). Basic concerns in test validations. In J. C. Alderson & A. Hughes (Eds.), *ELT documents III- Issues in language* (pp 135 – 151). London: The British Council.

Pelser, M. K. (2009). *The concurrent validity of learning potential and psychomotor ability measures for the selection of Haul Truck Operations in an open pit mine* (Unpublished master's thesis). University of South Africa (UNISA), South Africa.

Pensavalle, C. A., & Solinas, G. (2013). The Rasch Model Analysis for understanding mathematics proficiency - A case study: Senior high school Sardinian students. *Creative Education*, *4*(12), 767-773.

Perry, W. G. (1970). *Forms of Intellectual and Ethical Development in the College Years: A scheme.* New York: Holt, Rinehart and Winston.

Piaget, J., & Cook, M. (1977). *The origins of intelligence in the child.* University of California: Penguin.

Polk, T.A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review, 102*(3), 533-566.

Pretorius, E. J. (2002). Reading ability and academic performance in South Africa: Are we fiddling while Rome is burning? *Language matters*, *33*, 169-196.

Radden, G. (2008). The cognitive approach to language. In J. Andor, B. Hollosy, T. Laczko, & P. Pelyvas (Eds.), *When grammar minds language and literature, Festschrift for Prof Bela Korponay on the occasion of his 80th birthday* (pp. 387-412). Debrecen: Institute of English & American Studies.

Rambiritch, A. (2012). Challenging Messick: Proposing a theoretical framework for understanding fundamental concepts in language testing. *Journal of Language Teaching*, *46*(2), 108-121.

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: MESA Press.

Ravand, H., & Firoozi, T. (2016). Examining construct validity of the Masters UEE using the Rasch Model and the six aspects of the Messick's Framework. *International Journal of Language Testing*, *6*(1), 1-18.

Republic of South Africa. (1995). *Labour Relations Act 66 of 1995.* Cape Town: South African Government.

Republic of South Africa. (1998a). *Employment Equity Act 55 of 1998.* Cape Town: South African Government.

Republic of South Africa. (1998b). *Employment of Equity Bill* (Government Gazette, 400(19370), 14 February 2014). Pretoria: Government Printer

Republic of South Africa. (2014). *Employment Equity Amendment Act 47 of 2013.* Cape Town: South African Government.

Resmick, L. B. (2003). Standards and tests: Keeping them aligned. *Research Points, 1*(1), 1-4.

Reusser, K. (2001). Co-constructivism in educational theory and practise. In N. J. Smelser, B. Baltes & F. E. Weinert (Eds.), *International Encyclopedia of the Social and Behavioural Sciences* (pp. 2058-2062). Oxford, United Kingdom: Pergamon/Elsevier Science

Reyes, J. M., & Johnson, D. M. (2010). Construct validity of the Spanish version of a state-mandated high-stakes test (TAKS). *Journal of Border Educational Research, 8*, 59-69.

Rieber, R. W. (1983). Noam Chomsky's views on the psychology of language and thought. In R. W. Rieber & G. Voyat (Eds.), *Dialogues on the psychology of language and thought: Conversations with Noam Chomsky, Charles Osgood, Jean Piaget, Ulric Neisser and Marcel Kinsbourn* (pp. 29-63). New York: Plenum Press.

Ritter, N. L. (2010). *Understanding a widely misunderstood statistic: Cronbach's α.* Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans.

Roomaney, R., & Koch, E. (2013). An item and construct bias analysis of two language versions of a verbal analogies scale. *South African Journal of Psychology, 43*(3), 314-326.

Sabri, S. (2013). Item analysis of student comprehensive test for research in teaching beginner string ensemble using model based teaching among music students in public universities. *International Journal of Education and Research, 1*(12), 1-13.

Santrock, J. W. (2010). *Life-span development. 13th edition.* New York: McGraw Hill.

Santrock, J. W. (2013). *Life-span development. 14th edition*. New York: McGraw Hill.

Saville-Troike, M. (1984). What really matters in second language learning for academic achievement? *TESOL Quarterly 18*(2), 199-249.

Schaap, P. (2011). The differential item functioning and structural equivalence of a non-verbal cognitive ability test for five language groups. *South African Journal of Industrial Psychology 37*(1), 137-152.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychology Methods*, *7*(2), 147-177.

Schaie, K. W. (2008). A lifespan developmental perspective of psychological aging. In K. Laidlaw & B. G. Knight (Eds.), *The Handbook of emotional disorders in late life: Assessment and treatment* (pp. 3-32). Oxford, United Kingdom: Oxford University Press.

Schaie, K. W. & Willis, S. L. (1993). Age Difference Patterns of Psychometric Intelligence in Adulthood: Generalizability Within and Across Ability Domains. *Psychology and Aging*, 8(1), 44-55.

Schaie, K. W. & Willis, S. L. (2000). A stage theory model of adult cognitive development revised. In R. Rubinstein, M. Moss, & M. Kleban (Eds.), *The many dimensions of aging: Essays in honor of M. Powell Lawton* (pp. 175-193). New York: Springer Publishing Company.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353.

Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to Structural Equation Modelling* (3rd Ed.). New York: Routledge.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammon (Ed.), *Review of research in education, 19*, 406-450. Washington, DC: AERA.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477-481.

Sireci, S. G., & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated adapted items. *Applied Measurement in Education, 13*(3), 229-248.

Sireci, S. G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement Issues and Practices, 25*(3), 27-34.

Sijtsma, K. (2012). Psychological measurement between physics and statistics. *Theory and Psychology, 22*(6), 786-809.

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*(3), 281-311.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*(2), 205-231.

Smith, E. V., & Smith, R. M. (2004). *Introduction to Rasch measurement. Theory, models and applications.* Maple Grove, Minnesota: Jam Press.

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*, 66-78.

Smith, E. V., Wakely, M. B., De Kruif, R. E. L., & Swartz, C. W. (2003). Optimizing rating scales for self-efficacy (and other) research. *Educational and Psychological Measurement, 63*(3), 369-391.

Sotelo-dynega, M., Ortiz, S. O., Flanagan, D. P., & Chaplin, W. F. (2013). English language proficiency and test performance: an evaluation of bilingual students with the Woodcock-Johnson III tests of cognitive abilities. *Psychology in the Schools, 50*(8), 781-797.

Spearman, C. (1904). “General Intelligence”, Objectively Determined and Measured. *American Journal of Psychology*, 15, 201-293.

Spolsky, B. (1975). *Language testing: art of science*. Paper read at the 4th International Congress of Applied Linguistics. In proceedings of the 4th International Congress of Applied Linguistics. Stuttgurt: Hochschul Verlag, 9 – 28.

Sternberg, R. (1988). *The Triarchic Mind: A New Theory of Human Intelligence*. New York: Viking.

Sternberg, R. J., Jarvin, L., & Grigorenko, E. L. (2011). *Explorations of giftedness*. United States of America: Cambridge University Press.

Strand, S. (2004). Consistency in reasoning test scores over time. *British Journal of Educational Psychology*, *74*, 617-631.

Strand, A., Deary, I. J., & Smith, P. (2006). Sex differences in cognitive abilities test scores: A UK national picture. *British Journal of Educational Psychology*, *76*, 463-480.

Streiner, D. L. (2010). Measure for measure: New development in measurement and item response theory. *Canadian Journal of Psychiatry*, *55*(3), 180-186.

Struik, M. (2011). *The translation and validation of the postpartum depression screening scale (PDSS): Towards improving screening for postpartum depression in English and Afrikaans speaking South African women* (Unpublished doctoral dissertation). University of Pretoria, Pretoria, South Africa.

Suhr, D. D. (2006). *Exploratory or confirmatory factor analysis?* Paper presented at the SAS Users Group International Conference SUGI 31, San Francisco, CA.

Suhr, D. D., & Shay, M. (2009). *Guidelines for reliability, confirmatory and exploratory analysis.* Paper presented at the SAS Global Forum, Washington, D.C., USA.

Taylor, T. R. (1994). A review of three approaches to cognitive assessment, and a proposed integrated approach based on a unifying theoretical framework. *South African Journal of Psychology, 24*, 1-16.

Terblanche, S. (2015). *IBM SPSS Amos: Structural Equation Modeling with Amos*. Personal Collection of OLRAC SPS, South Africa (Johannesburg).

Thorndike, E. L. (1927). *The Measurement of Intelligence*. New York: Teachers College (Columbia University).

Thorpe, G. L., & Favia, A. (2012). Data analysis using Item Response Theory methodology: An introduction to selected programs and applications. *Psychology Faculty Scholarship, Paper 20,* 1-20. Retrieved from http://digitalcommons.library.umaine.edu/psy_facpub/20.

Thurstone, L. L. (1938). *Primary Mental Abilities*. Chicago: University of Chicago Press

Tomu, H. (2013). The role played by the Health Professions of South Africa (HPCSA) Ethical code of conduct and the Employment Equity Act (EEA) in regulating professional, legal and ethical conduct of psychologists in South Africa. *International Journal of Academic Research in Economics and Management Sciences, 2*(1), 59-66.

Tseng, W. S. (2001). *Handbook of cultural psychiatry*. New York, NY: Academic Press.

Valette, R. (1967). *Modern language testing: a handbook*. New York: Harcourt Brace World.

Van den Broek, P., & Gustafson, M. (1999). Comprehension and memory for texts. Three generations of reading research. In S. R. Goldman, A. C. Graesser, & P. Van den Broek (Eds.), *Narrative comprehension, causality and coherence: Essays in honor of Tom Trabasso* (pp. 15-34) Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Van der Elst, W., Reed, H., & Jolles, J. (2013). The logical grammatical structures test: Psychometric properties and normative data in Dutch-speaking children and adolescents. *The Clinical Neuropsychologist, 27*(3), 396-409.

Van der Heijden, P. G. M., Van Buuren, S., Radder, J., & Verrips, E. (2003). Unidimensionality and reliability under Mokken scaling of the Dutch language version of the SF36. *Quality of Life Research, 12*, 189-198.

Van der Pool, M., & Catano, V. M. (2008). Comparing the performance of native north Americans and predominately white military recruits on verbal and non-verbal measures of cognitive ability. *International Journal of Selection and Assessment, 16*(3), 239-248.

Van de Vijver, F. J. R., & Rothmann, S. (2004). Assessment in multicultural groups. The South African case. *South African Journal of Industrial Psychology, 30*(4), 1-7.

Van de Vijver, F. & Tanzer, N. K. (2004). Bias and Equivalence in Cross-Cultural Assessment: an overview. *Revue Europeenne de psychologie appliquee*, 54, 119-135.

Van der Walt, J. L., & Steyn, H. S. (2008). The validation of language tests. *Stellenbosch Papers in Linguistics, 38*, 191-204.

Verhelst, N. D., & Jansen, M. G. H. (1992). A logistic model for time limit tests. *Measurement and Research Department Reports, 92*, 1-28.

Vernon, P. E. (1950). *The Structure of Human Abilities*. London, England: Methuen.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Wankat, P. C., & Oreovicz, F. S. (1993). *Teaching Engineering*. New York: McGraw Hill.

Weir, C. J. (2005). *Language testing and validation*. Basingstoke: Palgrave Macmillan

Whitfield, K. E., Allaire, J. C., Gamaldo, A. A., & Bichsel, J. (2010). Factor structure of cognitive ability measures in older African Americans. *Journal of Cross-Cultural Gerontology, 25(3)*, 271-284.

Willis, S. L. & Schaie, K. W. (2006). A co constructionist View of the Thrid Age: The Case of Cognition. *Annual Review of Gerontology and Geriatrics*, 26, 131-151.

Wolming, S., & Wikstrom, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy and Practice, 17* (2), 117-132.

Woodcock, R. W., & Muñoz-Sandoval, A. F. (2005). *WMLS-R WMLS R.Xml learning, memory and language assessment.* Itasca, IL: Riverside Publishing.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practise (Winter)*, 33-45.

Yong, A. C., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology, 9*(2), 79-94.

Yu, C. H. (2013). *A simple guide to the Item Response Theory (IRT) and Rasch Modelling.* Retrieved from http://www.creativewisdom.com/computer/sas/IRT.pdf

Yun, J., & Ulrich, D. A. (2002). Estimating measurement validity: A tutorial. *Adapted Physical Activity Quarterly*, *19*, 32-47.

Zhang, Z., & Takane, Y. (2010). Statistics: Multidimensional scaling. In E. Baker, B. McGraw, & P. P. Petersen, (Eds.), *International encyclopedia of education* (pp. 304-311). Oxford, UK: Elsevier.

Zumbo, B. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing*, *20*, 136-147.

APPENDIX A: ENGLISH COMPREHENSION TEST (ECT)

ENGLISH COMPREHENSION TEST (ECT)

Instructions:

1. **READ the instructions for the ECT carefully before answering the test.**
2. **The ECT consists of SECTION A and SECTION B. Please make sure that you answer all the questions CORRECTLY on the ANSWER SHEET provided.**
3. **Answer ALL questions.**
4. **When uncertain, choose the MOST CORRECT answer.**
5. **There is NO TIME LIMIT imposed, but work at an EVEN PACE.**
6. **Do not WRITE OR MAKE ANY MARKINGS on this question booklet.**
7. **Try your best and Good luck!**

A MILITARY STORY

1. Warrant Officer (WO) Moses who worked in one of the military units was sent to a foreign country to attend to some work. On the first leg of the journey he landed and taxied to his chosen parking space. Everything went very well but when he looked out of the window of the aircraft, he saw that aircrafts from various companies were parked with their wings overlapping one another and he became nervous.

2. He was told that refuelling was going to take place and that all the passengers had to disembark and stand on the flight line towards the back of the aircraft ("Odd", he thought).

3. It was a very humid day when they landed and the traffic was quite heavy. While they watched the aircraft taking off and landing, some of the passengers smoked and stretched their legs after the three-hour flight.

4. He was amazed at all the litter and rubbish lying around. There were empty hydraulic oil cans, scraps of paper, bolts from aircraft, pieces of luggage, metal clips, empty food containers, shell casings, a lot of aircraft odds and ends and wait for it, live AK 47 cartridges.

5. Moreover, they were told not to wander too far away from the airstrip, as there were landmines in the area. All the rubbish and mines lay within an approximate radius of 20 m from the large turbine engines that were starting up and the aircraft taxiing past. What a recipe for disaster!

6. Adding to the risks of this potential disaster, there were many people just wandering around the airport and across the main runway. He realised that they were in a place where the first priority was to get on with the job, load and go.

7. He wondered who was responsible for aviation safety at this particular international airport. He realised that unsafe practices are very obvious to anyone who has completed the basic Aviation Safety course. When he got to the headquarters, he thanked Lt Col Hoekstra for equipping him with this knowledge as he now knows how airports should be run in order to ensure safety.

(Adapted from a letter entitled,
"To be or not to be aviation safe", Nyala, 2004)

SECTION A: COMPREHENSION

- ❖ **Please make sure that you have read the passage before answering the questions for Section A.**
- ❖ **Please answer all the questions on the answer sheet provided.**

1. Which statement is TRUE according to the information given in the passage?

A. People were walking on the runway of the airport.
B. People wandered into the area where the landmines were.
C. People littered and left rubbish at the airport.
D. People had the first priority of getting on with the job.

2. Which statement is TRUE according to the information given in the passage?

A. The passengers got out of the aircraft because they wanted to see the back of the aircraft during refuelling.
B. The passengers got out of the aircraft because they wanted to stand by the flight line during refuelling.
C. The passengers got out of the aircraft because they could not be in the aircraft during refuelling.
D. The passengers got out of the aircraft because they felt odd and nervous during refuelling.

3. Which statement is FALSE according to the information given in the passage?

A. WO Moses knows how to run airports in order to ensure safety.
B. WO Moses knew who was responsible for aviation safety at the airport.
C. Lt Col Hoekstra helped WO Moses with the Aviation Safety course.
D. WO Moses was not sure who was responsible for the safety at the airport.

4. Which statement is TRUE according to the information given in the passage?

A. WO Moses went to a different airport.
B. WO Moses went to a new airport.
C. WO Moses went to inspect airports after the Aviation Safety course.
D. WO Moses went to a different country.

5. Which statement is FALSE according to the information given in the passage?

A. The airport was thought of as a potential disaster.
B. The litter and rubbish made the airport a potential disaster.
C. The people wandering around the airport runway made it a potential disaster.
D. The airport was a disaster area.

6. Which statement is FALSE according to the information given in the passage?

A. WO Moses saw that aircrafts from South African companies were parked with their wings overlapping.
B. WO Moses did not feel good about what he saw from the aircraft window.
C. WO Moses saw more than one aircraft from his window.
D. WO Moses saw that the wings of the different aircrafts were next to each other.

7. Indicate whether the following statements are either a <u>FACT</u> (these are true statements) or an <u>OPINION</u> (these are judgments or beliefs).

7.1 WO Moses was thanked by Lt Col Hoekstra for helping him with the Aviation Safety course.

7.2 WO Moses recognized the unsafe practices because he was a pilot and did the Aviation Safety course.

7.3 There was a lot of rubbish and litter lying around on the airport strip.

7.4 There were landmines in the area that the passengers had landed in.

7.5 The passengers smoked when they landed because the flight was long and made them tired.

8. Which statement below summarizes what '<u>the first leg of the journey'</u> (par. 1) means?

A. The beginning of every flight overseas.
B. The arrival at the journey.
C. The first phase of a trip.
D. The first refuelling part of the journey.

9. Which statement below summarizes the quote, <u>"a recipe for disaster"</u> the best?

A. WO Moses had the ingredients to make the disaster.
B. The litter and rubbish in the area made it unsafe.
C. The airport was a disaster.
D. WO Moses thought that the airport was untidy.

10. Based on the passage, what <u>ADVANTAGES</u> listed below did the Aviation Safety course have for WO Moses?

A. He was aware of the aviation safety guidelines and did not like the litter that he saw at the airport.
B. He was aware of the unsafe aviation practices and he wanted to speak to the person responsible at the airport for the rubbish and litter.
C. He was more knowledgeable about aviation safety and he wanted to become a safe pilot.
D. He was aware of the aviation safety guidelines and he knew how airports should be run in order to create a safe environment.

11. Based on the passage, what is the <u>DISADVANTAGE</u> listed below of unsafe aviation

practices?

A. The litter and rubbish lying around the airport makes it ugly and less people will use the airport.
B. The litter and rubbish lying around the airport can be hazardous for the airport.
C. The airport is unsafe for important people and foreigners visiting the country.
D. The airport is popular for the litter and rubbish lying around.

12. Which word best describes WO Moses's response to Lt Col Hoekstra?

A. Satisfied
B. Content
C. Thankful
D. Helpful

13. What was the purpose of the article?

A. Aviation safety affects your reputation in the military.
B. Litter and rubbish lying around makes the airport ugly.
C. The pilots at the airport caused the rubbish and litter to lie around.
D. The importance of being aviation safe.

SECTION B: LANGUAGE

- ❖ **Section B does not involve the comprehension piece (A Military Story).**
- ❖ **Please answer all the questions on the answer sheet provided.**

14. Choose the CORRECT WORD/TENSE in the following statements.

14.1. The Army General wanted to know **(A. which, B. whom, C. who, D. whose)** had gone to fetch the other soldiers.

14.2. (A. This, B. These, C. That, D. There) pilots were doing different stunts with their aircrafts when birds flew beside them.

14.3. (A. They, B. Whose, C. Who, D. Which) knew which army was about to attack and wanted to prepare the other soldiers.

14.4. WO Moses and his colleagues **(A. is, B. were, C. was, D. has)** deployed to Central Africa last year.

14.5. The battle was called CODE-D, because of the brave soldier **(A. that, B. who, C. whose, D. which)** life was innocently taken in battle.

14.6. The documents which were given to the General **(A. was, B. were, C. is, D. has been)** under inspection.

15. Choose the best word from the list of words provided below which has the SAME MEANING (synonym) as the following words.

15.1 Delegate.
A. Gateway.
B. Fragile.
C. Entrust.
D. Question.

15.2 Solicit.
A. Describe.
B. Ask for.
C. Explanation.
D. To exclaim.

15.3 Inferior.
A. Higher.
B. Undermine.
C. Lower.
D. Smaller.

15.4 Inaugural.
A. Ceremony.
B. Majestic.
C. Impartial.
D. Introductory.

15.5 Vague.
A. Unclear.
B. Clear.
C. Safe.
D. Quick.

16. Choose the CORRECT FORM OF THE WORD in the following sentences.

16.1. The army **(A. man, B. men)** were very tired after their war battle exercise in Pretoria.

16.2. The pilots were flying across the midlands when they saw **(A. people, B. person)** waving to them.

16.3. The navy men dived into the sea and injured their **(A. foot, B. feet)** against the hidden rocks.

16.4. While the medics were on duty for the hospital strikes, they saw seriously injured **(A. children, B. child,)** in the ward.

16.5. While the new navy divers were training at Simons Town, they saw a school of **(A. fishes, B. fish)** in the sea.

17. Choose the best word from the list of words provided below which has an <u>OPPOSITE MEANING (antonym)</u> to the following words.

17.1. <u>Disembark</u>:
A. Change positions
B. Get off
C. Get on
D. Land

17.2. <u>Hazardous</u>:
A. Dangerous
B. Safe
C. Risky
D. Not clear

17.3. <u>Desperate:</u>
A. Happy
B. Nervous
C. Worried
D. Peaceful

17.4. <u>Foreigner:</u>
A. Local
B. Stranger
C. Xenophobia
D. Visitor

17.5. <u>Approximately:</u>
A. Estimated
B. Almost perfect
C. Accurate
D. Roughly

18. REARRANGE THE WORDS below in order TO MAKE A SENTENCE. <u>ALL the words must be used in the sentence</u>.

18.1. all users of the clean airports need to be kept at all times to ensure safety.

18.2. time calculated involves a job making decisions about pilot's and space.

18.3. their colleagues are integrity by people with greatly respected.

18.4. not so legitimate unless uniform members doing wear must for a reason all there is.

¥

†

MILITARY PSYCHOLOGICAL INSTITUTE

SIDE 1

ECT

ID NO.													

	0	1	2	3	4	5	6	7	8	9
GROUP	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○
CANDIDATE NO	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○
AGE	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○

	0	1	2	3	4	5	6	7	8	9
ID NO	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○
	○	○	○	○	○	○	○	○	○	○

	AFRIKAANS	ENGLISH	ISINDEBELE	N. SOTHO	S. SOTHO	SISWATI	XITSONGA	SETSWANA	TSHIVENDA	ISIXHOSA	ISIZULU	OTHER
SCHOOL 1ST LANG	○	○	○	○	○	○	○	○	○	○	○	○
SCHOOL 2ND LANG	○	○	○	○	○	○	○	○	○	○	○	○
HOME LANG	○	○	○	○	○	○	○	○	○	○	○	○

SURNAME																						
INITIALS																						
SCHOOL																						
TOWN																						
PROVINCE																						

SECTION A: COMPREHENSION

	A	B	C	D
1	○	○	○	○
2	○	○	○	○
3	○	○	○	○
4	○	○	○	○
5	○	○	○	○
6	○	○	○	○

	FACT	OPINION
7.1	○	○
7.2	○	○
7.3	○	○
7.4	○	○
7.5	○	○

	A	B	C	D
8	○	○	○	○
9	○	○	○	○
10	○	○	○	○
11	○	○	○	○
12	○	○	○	○
13	○	○	○	○

SECTION B: LANGUAGE

	A	B	C	D
14.1	○	○	○	○
14.2	○	○	○	○
14.3	○	○	○	○
14.4	○	○	○	○
14.5	○	○	○	○
14.6	○	○	○	○
15.1	○	○	○	○
15.2	○	○	○	○
15.3	○	○	○	○
15.4	○	○	○	○
15.5	○	○	○	○

	A	B
16.1	○	○
16.2	○	○
16.3	○	○
16.4	○	○
16.5	○	○

	A	B	C	D
17.1	○	○	○	○
17.2	○	○	○	○
17.3	○	○	○	○
17.4	○	○	○	○
17.5	○	○	○	○

† ¥

18.1……………………………………………………………………………………………………

……

……

18.2……………………………………………………………………………………………………

……

……

18.3……………………………………………………………………………………………………

……

……

18.4……………………………………………………………………………………………………

……

……